DCS-Mamba: Linear-Complexity Spatiotemporal Modeling with Semantic Modulation for High-Resolution Remote Sensing

Jiachen Yuan[1] [0009-0000-8763-0440], Wenzheng Huang[1] [0009-0004-0969-7331], Yihan Huang[1] [0009-0007-4594-9753], Zihan Ye[1] [0009-0002-8529-285X], Yanqin Shi[1] [0009-0005-5429-2496], Xianyi Yang[2] [0000-0003-4714-9112], Ning Xin[1][0009-0003-7718-6552]*, Md Maruf Hasan[2][0000-0001-9809-9376]**

1 School of AI and Advanced Computing, XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China
[2] Center for Marine Meteorology and Climate Change, State Key Laboratory of Marine Environmental Science, College of Ocean and Earth Sciences, Xiamen University, Xiamen, China
Jiachen.Yuan23@student.xjtlu.edu.cn

Abstract. High-resolution spatiotemporal forecasting of atmospheric pollutants remains challenging because it requires long-horizon modeling over dense remote-sensing grids while staying responsive to abrupt meteorological events. Existing convolutional and recurrent predictors preserve local patterns but struggle with long-range dependencies, whereas attention-based models are accurate yet expensive on large spatial fields. We propose DCS-Mamba, a decomposed context-aware framework that combines trend-seasonal distribution alignment, selective state-space scanning, and text-guided adaptive group normalization. The model separates slowly varying background evolution from short- term residual dynamics, scans pixel-wise temporal tokens with linear complexity, and injects meteorological semantics into latent physical features. Experiments on 2020–2024 real-world air-quality data and two additional geographical regimes show that DCS-Mamba outperforms strong video-prediction, Transformer, and foundation-model baselines. On the 2024 test set, it reduces Physical MSE by 9.3% and Normalized MSE by 20.6% compared with SimVP, while maintaining sharper spatial reconstructions.

Keywords: Remote Sensing · Spatiotemporal Prediction · Multimodal Learning · State Space Models

1 Introduction

Spatiotemporal forecasting of atmospheric pollutants, including $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3, is important for public health, urban management, and environmental policy. Unlike ordinary time-series prediction, pixel-level forecasting over high-

* Corresponding author.
** Corresponding author.

resolution remote-sensing grids requires the model to capture complex atmospheric transport, local emissions, seasonal changes, and terrain-dependent dispersion at the same time [1, 2, 3]. Higher spatial precision improves the ability to identify local pollution hotspots, but it also increases memory and computation costs for modern deep models [4].

Existing approaches still face three limitations. First, CNN- and RNN-based video prediction models, including ConvLSTM, SimVP, and MMVP, are effective in preserving local structures but are fundamentally constrained by local receptive fields and recurrent information bottlenecks over long horizons [5, 6, 7]. Second, spatiotemporal Transformer models such as Earthformer and PredFormer improve global dependency modeling, yet their attention operations are costly when the spatial field is dense and the temporal horizon is long [8, 9, 10]. Third, recent foundation or time-series models such as ClimaX, Time-LLM, and PatchTST have shown promise for environmental or temporal prediction, but they mainly rely on numerical sequences and often underuse semantic descriptions of meteorological events [11, 12, 13, 14]. Such linguistic information can be crucial when sudden weather-driven shifts occur.

To address these limitations, we propose DCS-Mamba, a Decomposed Context-aware Spatiotemporal Mamba framework for pixel-level earth observation. DCS-Mamba uses selective state-space modeling to capture long-range temporal dependencies efficiently [15]. It further integrates Reversible Instance Normalization and trend-seasonal decomposition to stabilize non-stationary pollutant dynamics [17, 18]. To bridge the semantic-physical gap, we introduce a Text-Guided Adaptive Group Normalization module that converts meteorological prompts into modulation parameters for latent physical features [19].

The main contributions are summarized as follows. First, we propose a Mamba-based high-resolution pollutant forecasting framework that preserves long-range spatiotemporal dependencies without constructing dense attention maps. Second, we design a semantic modulation mechanism that injects meteorological context into physical feature transitions through adaptive normalization. Third, we combine RevIN with trend-seasonal decomposition to explicitly handle severe distribution shifts in atmospheric data. Fourth, experiments on the 2024 test set and two geographically distinct datasets demonstrate that DCS-Mamba consistently improves physical and normalized metrics while preserving fine-grained spatial structures.

2 Related Work

Spatiotemporal air quality forecasting has gradually shifted from traditional statistical models to deep learning methods. Early studies commonly adopted CNN-LSTM hybrids [5] to jointly capture spatial features and temporal dynamics, but these models are limited by local receptive fields and high computational costs over long forecasting horizons. To enhance spatial modeling, recent methods have incorporated multisource meteorological and satellite data through advanced convolutions and graph networks [4, 14]. However, they still struggle with non-stationary atmospheric variations. U-Net [20] remains widely used for preserving multi-scale spatial resolution, yet its

reliance on 3D convolutions or Transformer-based temporal modules often introduces substantial overhead on high-resolution remote sensing grids.

State Space Models (SSMs) have recently attracted attention as efficient alternatives to self-attention. Mamba [15] introduces a selective scan mechanism that supports long-range sequence modeling with linear complexity and a global receptive field. Its success has also been extended to visual modeling [16], showing promise for remote sensing applications [16]. Nevertheless, standard SSM-based methods mainly operate on numerical features and lack explicit semantic modulation for sudden environmental changes caused by external meteorological events.

Non-stationarity remains another key challenge in climate and air quality forecasting. Autoformer [18] addresses this issue through trend-seasonal decomposition, which helps stabilize temporal prediction. Meanwhile, multimodal learning has shown potential in urban computing by combining numerical observations with contextual information [14]. Recent foundation models such as Time-LLM [12] further explore language-model reprogramming for time-series forecasting, but they mainly focus on one-dimensional sequences and are less suitable for dense two-dimensional physical fields. In contrast, DCS-Mamba combines non-stationary trend decomposition [18] with Text-Guided AdaGN to bridge semantic and physical representations, enabling efficient high-resolution forecasting against strong baselines such as SimVP [6] and Time-LLM [12].

3 Methodology

3.1 Data Preprocessing

As illustrated in Fig. 1, we adopt a preprocessing pipeline to ensure spatial consistency, physical validity, and temporal continuity. The raw satellite observations are first realigned to a uniform grid of resolution 192×256 using bilinear interpolation. This resolution provides a practical balance between local spatial fidelity and GPU memory cost. Owing to the complex topography of the study region, approximately one third of the bounding boxes correspond to unmonitored areas, including large water bodies or sparsely inhabited mountainous regions. These invalid regions are filtered by a binary spatial mask $\mathcal{M} \in \{0,1\}^{H \times W}$. Pixels with $\mathcal{M}_{h,w} = 0$ are excluded from both forward prediction and objective computation, preventing spurious gradients over physically invalid locations. Missing temporal values are filled by linear interpolation within the historical observation window, while spatial holes are reconstructed by inverse distance weighting from neighboring valid pixels. All imputations are performed without using future ground-truth observations. Semantic prompts are generated by a language model from historical pollutant statistics and meteorological context within the observed input window only, ensuring that the forecasting protocol remains causal.

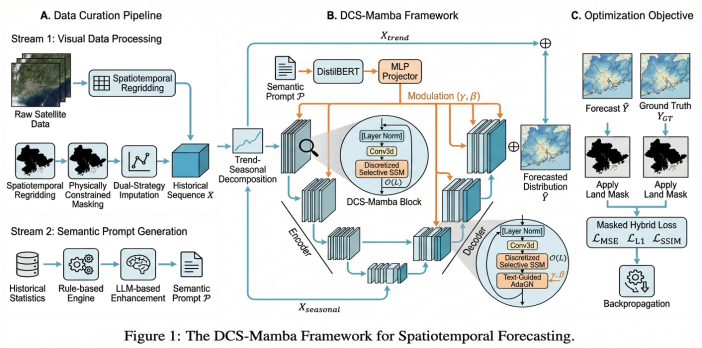


Figure 1: The DCS-Mamba Framework for Spatiotemporal Forecasting.

Fig. 1. Overall architecture of the proposed DCS-Mamba framework. The pipeline performs physically valid data curation, decomposes non-stationary inputs into trend and seasonal streams, models long-range dependencies through selective scans, and injects semantic context through Text-Guided AdaGN before masked objective optimization.

3.2 DCS-Mamba Architecture

The proposed DCS-Mamba follows a U-shaped encoder-decoder architecture enhanced by a dual-branch decomposition strategy. To address non-stationarity, long-range dependency modeling, and cross-modal alignment in atmospheric forecasting, the framework contains three coupled components: non-stationary distribution alignment, linear-complexity spatiotemporal selective scanning, and text-guided adaptive modulation. The trend branch captures slowly varying background evolution, while the seasonal branch preserves high-frequency residual dynamics that are forwarded to the Mamba encoder. Text-guided AdaGN is inserted into the decoder to adjust latent physical features according to historical meteorological semantics.

3.3 Disentangled Modeling of Non-Stationary Dynamics

Atmospheric observations often exhibit temporal distribution shifts induced by seasonal variation, diurnal cycles, and changing meteorological regimes. To alleviate this non-stationarity, we combine Reversible Instance Normalization with trend-seasonal decomposition. Given a raw input sequence $\mathcal{X} \in \mathbb{R}^{B\times T\times C\times H\times W}$, we normalize each sample along the temporal dimension:

$$\mathcal{X}_{\text{norm}} = \frac{\mathcal{X}-\mu}{\sigma+\epsilon}, \quad \mu = \text{Mean}(\mathcal{X}, \text{dim} = T), \quad \sigma = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\mathcal{X}_{:,t,:,:,:}-\mu\right)^2}. \tag{1}$$

Here $\mu, \sigma \in \mathbb{R}^{B\times 1\times C\times H\times W}$ denote the instance-wise temporal mean and standard deviation, and ϵ is a stabilizing constant.

After normalization, $\mathcal{X}_{\text{norm}}$ is decomposed into a low-frequency trend component $\mathcal{X}_{\text{trend}}$ and a high-frequency seasonal component $\mathcal{X}_{\text{season}}$ using a moving-average operator with window size K:

$$\mathcal{X}_{\text{trend}}^{(i)} = \frac{1}{K} \sum_{j=i-\left\lfloor \frac{K}{2} \right\rfloor}^{i+\left\lfloor \frac{K}{2} \right\rfloor} \text{Pad}(\mathcal{X}_{\text{norm}})^{(j)}, \quad \mathcal{X}_{\text{season}} = \mathcal{X}_{\text{norm}} - \mathcal{X}_{\text{trend}}. \tag{2}$$

The trend branch captures slowly varying background evolution, whereas the seasonal branch preserves higher-frequency residual dynamics. After decoding, the normalized prediction $\hat{\mathcal{Y}}_{\text{norm}}$ is restored to the original physical scale through the inverse RevIN transform:

$$\hat{\mathcal{Y}} = \hat{\mathcal{Y}}_{\text{norm}} \odot (\sigma + \epsilon) + \mu. \tag{3}$$

This reversible normalization decouples changing marginal statistics from the underlying forecasting dynamics while preserving physical units in the output.

3.4 Linear-Complexity Spatiotemporal State Space Model

The encoder replaces quadratic self-attention with the Selective State Space Model. Instead of explicitly modeling pairwise token interactions, the SSM captures long-range temporal dependencies through latent state evolution. A continuous-time SSM maps a one-dimensional input $u(t)$ to an output $y(t)$ through a latent state $h(t)$:

$$h'(t) = \boldsymbol{A}h(t) + \boldsymbol{B}u(t), \quad y(t) = \boldsymbol{C}h(t). \tag{4}$$

To implement this system on digital hardware, we discretize Eq. (4) using zero-order hold:

$$\overline{\boldsymbol{A}} = \exp(\Delta \boldsymbol{A}), \quad \overline{\boldsymbol{B}} = (\Delta \boldsymbol{A})^{-1}(\exp(\Delta \boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta \boldsymbol{B}. \tag{5}$$

Standard SSMs are linear time-invariant. We adopt the S6 parameterization, in which the input-dependent quantities $\boldsymbol{B}_t$, $\boldsymbol{C}_t$, and Δ_t are generated dynamically:

$$\boldsymbol{B}_t, \boldsymbol{C}_t, \Delta_t = \text{LinearProj}(u_t). \tag{6}$$

Given the seasonal feature map $\mathcal{X}_{\text{season}}$, we first apply a spatial embedding and then flatten spatial positions into independent temporal tokens. Let $u_t := \mathcal{Z}_{\text{flat}}[t]$ denote the flattened seasonal token at time step t. The resulting discrete recurrence is

$$h_t = \overline{\boldsymbol{A}}_t h_{t-1} + \overline{\boldsymbol{B}}_t u_t, \quad y_t = \boldsymbol{C}_t h_t. \tag{7}$$

By factorizing spatial and temporal processing, the model performs temporal scans over flattened spatial positions while preserving linear complexity in sequence length. Ignoring constant channel projection terms, the dominant complexity is

$$\Omega_{\text{Mamba}} = \mathcal{O}(HW \cdot T). \tag{8}$$

Table 1. Computational complexity comparison. M denotes the local attention window size in Swin Transformer.

Architecture	Spatial Complexity	Temporal Complexity
Full Self-Attention	$\mathcal{O}(T \cdot (HW)^2)$	$\mathcal{O}(HW \cdot T^2)$
Swin Transformer	$\mathcal{O}(T \cdot HW \cdot M^2)$	$\mathcal{O}(HW \cdot T^2)$
DCS-Mamba	$\mathcal{O}(\boldsymbol{T} \cdot \boldsymbol{HW})$	$\mathcal{O}(\boldsymbol{HW} \cdot \boldsymbol{T})$

3.5 Semantic-Physical Fusion via Adaptive Modulation

To reduce the modality gap between textual priors and latent physical features, we introduce Text-Guided Adaptive Group Normalization. Given a natural-language prompt $\mathcal{P}$ derived from historical atmospheric summaries, a pre-trained DistilBERT encoder followed by mean pooling obtains a fixed-length semantic embedding:

$$\boldsymbol{e}_{\text{text}} = \text{MeanPool}\big(\text{DistilBERT}(\mathcal{P})\big). \tag{9}$$

This embedding is projected by a multi-layer perceptron to generate modulation parameters

$$\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{B \times G}, \tag{10}$$

where G denotes the number of channel groups. Let $\boldsymbol{Z} \in \mathbb{R}^{B \times T \times C \times H \times W}$ be the SSM feature map and let $g(c)$ denote the group index of channel c. We first compute group-wise statistics

$$\mu_g, \sigma_g \in \mathbb{R}^{B \times T \times G \times 1 \times 1}, \tag{11}$$

and then modulate the normalized features as

$$\widetilde{\boldsymbol{Z}}_{:,:,c,:,:} = \frac{\boldsymbol{Z}_{:,:,c,:,:} - \mu_{g(c)}}{\sqrt{\sigma_{g(c)}^2 + \epsilon}} \odot \left(1 + \boldsymbol{\gamma}_{:,g(c)}\right) + \boldsymbol{\beta}_{:,g(c)}. \tag{12}$$

The modulation parameters are broadcast along temporal and spatial dimensions. This design injects semantic context into latent feature dynamics while preserving the normalization structure used by the visual stream.

3.6 Objective Optimization

As shown in Fig. 1, the model is optimized with a masked hybrid loss over valid land pixels:

$$\Omega = \left\{ (h, w) \mid \mathcal{M}_{h,w} = 1 \right\}. \tag{13}$$

Let $\delta_{b,t,c,h,w} = \hat{\mathcal{Y}}_{b,t,c,h,w} - \mathcal{Y}_{b,t,c,h,w}$ denote the prediction residual. The masked squared and absolute losses are defined as

$$\mathcal{L}_2 = \frac{1}{BTC|\Omega|}\sum_{b,t,c}\sum_{(h,w)\in\Omega}\delta_{b,t,c,h,w}^2, \quad \mathcal{L}_1 = \frac{1}{BTC|\Omega|}\sum_{b,t,c}\sum_{(h,w)\in\Omega}\left|\delta_{b,t,c,h,w}\right|. \tag{14}$$

The overall training objective is

$$\mathcal{L} = \lambda_1\mathcal{L}_2 + \lambda_2\mathcal{L}_1 + \lambda_3\mathcal{L}_{\mathrm{SSIM}}, \tag{15}$$

where $\mathcal{L}_{\mathrm{SSIM}}$ denotes the Structural Similarity Index loss computed on valid spatial regions, encouraging preservation of local pollutant distribution structures.

4 Experiments

4.1 Multivariate Spatiotemporal Data Analysis

We conduct experiments on a comprehensive real-world atmospheric dataset sourced from the national meteorological bureau. The data spans 2020–2024, divided into training data from 2020–2022, validation data from 2023, and test data from 2024. Table 2 summarizes the statistical distribution of six pollutants. O_3 and PM_{10} exhibit high variance and large maxima, making extreme-value forecasting particularly challenging.

Table 2. Statistics of the collected real-world air-quality dataset. CO is reported in milligrams per cubic meter; the other indicators are reported in micrograms per cubic meter.

Indicator	Max	25% Q.	Mean	Median
$PM_{2.5}$	210.20	12.36	20.93	18.74
PM_{10}	352.46	23.86	36.38	32.94
SO_2	44.71	6.12	7.05	6.79
NO_2	228.17	10.08	16.08	13.63
CO	6.66	0.54	0.61	0.60
O_3	326.56	65.33	88.21	84.17

A key feature of our setup is the integration of semantic context. As shown in Fig. 2, air-quality fluctuations are strongly correlated with meteorological events, such as improved air quality during typhoon season and deterioration during winter monsoon periods. Text-Guided AdaGN is designed to leverage these event-level alignments without using any future target observations.

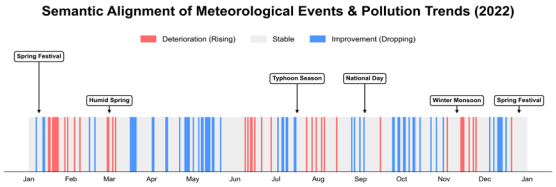


Fig. 2. Semantic alignment of meteorological events with pollution trends in 2022. Red indicates deterioration-related events, and blue indicates improvement-related events.

Fig. 3 displays the spatial distribution of the pollutants on a representative day. The high-resolution grid captures fine-grained hotspots, while masked water or unmonitored areas are excluded to ensure physical consistency. The input horizon is 45 days and the prediction horizon is 30 days. AdamW is used for optimization with the three loss weights set to 1.0, 1.0, and 0.5, respectively.

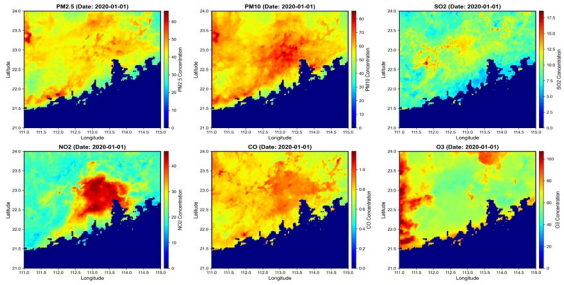


Fig. 3. Spatial distribution of six pollutants on Jan. 1, 2020. Dark blue regions represent unmonitored areas excluded by the physical mask.

4.2 Quantitative Comparison

We compare DCS-Mamba with eight representative baselines from video prediction, foundation/time-series modeling, and spatiotemporal Transformer families. Both physical metrics in real-world units and normalized metrics in z-score space are reported. As shown in Table~3, DCS-Mamba consistently achieves the best performance on the 2024 test set and two generalization datasets.

On the main test set, DCS-Mamba outperforms the strongest video prediction baseline, SimVP, reducing physical MSE from 180.33 to 163.56 and normalized MSE from 0.34 to 0.27. This indicates that the combination of Mamba-based long-range modeling and Text-Guided AdaGN captures spatiotemporal air-quality patterns more effectively than conventional CNN-based approaches. In contrast, general-purpose models such as Time-LLM and PatchTST produce much higher errors, suggesting that one-dimensional temporal models are less suitable for dense pixel-wise regional forecasting.

DCS-Mamba also generalizes well across different geographical regimes. On CI-AQ, it achieves the lowest physical MAPE of 14.43\%, while on MT-AQ, it reduces normalized MSE to 0.28 compared with 0.32 for SimVP. These results demonstrate its

robustness under both coastal industrial and mountainous terrain conditions, supporting its applicability to broader earth-observation scenarios.

Table 3. Quantitative comparison across the main 2024 test set and two diverse geographical regimes. Values for CI-AQ and MT-AQ represent average metrics across all six pollutant variables. Best results are bolded and colored in red; second best is colored in blue.

Method	Physical Metrics			Normalized Metrics		
	MSE	MAE	MAPE	MSE	MAE	MAPE
Main 2024 Test Set						
Earthformer (NeurIPS 2022)	201.01	7.36	23.42%	0.43	0.48	26.77%
SimVP (CVPR 2022)	180.33	6.72	19.87%	0.34	0.42	22.98%
SwinLSTM (ICCV 2023)	257.17	7.73	21.80%	0.43	0.45	25.31%
PatchTST (ICLR 2023)	452.34	11.40	33.55%	1.12	0.84	42.89%
ClimaX (ICML 2023)	234.32	7.76	23.28%	0.44	0.47	26.91%
MMVP (ICCV 2023)	313.60	8.96	25.80%	0.61	0.56	30.37%
Time-LLM (ICLR 2024)	355.18	9.98	30.23%	0.91	0.75	38.14%
PredFormer (TMLR 2026)	214.94	7.78	24.50%	0.49	0.50	28.43%
DCS-Mamba (Ours)	**163.56**	**6.29**	**18.77%**	**0.27**	**0.37**	**21.57%**
CI-AQ Dataset (Coastal Industrial Regimes)						
Earthformer (NeurIPS 2022)	486.07	11.86	28.91%	0.79	0.58	32.00%
SimVP (CVPR 2022)	386.36	10.39	24.62%	0.69	0.55	30.16%
SwinLSTM (ICCV 2023)	459.23	10.59	25.84%	0.62	0.51	31.74%
PatchTST (ICLR 2023)	538.09	12.89	32.59%	1.09	0.74	40.88%
ClimaX (ICML 2023)	309.87	**8.11**	18.38%	0.45	0.46	27.61%
MMVP (ICCV 2023)	508.49	11.96	28.46%	0.85	0.63	34.05%
Time-LLM (ICLR 2024)	444.92	11.42	28.50%	0.89	0.67	36.95%
PredFormer (TMLR 2026)	385.82	10.45	25.23%	0.71	0.56	31.86%
DCS-Mamba (Ours)	**291.74**	8.46	**14.43%**	**0.42**	**0.42**	**24.90%**
MT-AQ Dataset (Mountainous Terrain Regimes)						
Earthformer (NeurIPS 2022)	191.19	7.11	21.32%	0.49	0.49	29.14%
SimVP (CVPR 2022)	139.72	5.87	16.92%	0.32	0.39	24.22%
SwinLSTM (ICCV 2023)	164.43	6.07	16.90%	0.39	0.41	26.16%
PatchTST (ICLR 2023)	208.21	7.86	25.04%	0.81	0.68	38.90%
ClimaX (ICML 2023)	161.19	6.45	18.80%	0.45	0.46	28.13%
MMVP (ICCV 2023)	202.36	7.28	20.92%	0.62	0.55	30.75%
Time-LLM (ICLR 2024)	207.51	7.39	21.95%	0.61	0.58	33.55%
PredFormer (TMLR 2026)	149.84	6.38	19.02%	0.49	0.49	28.72%
DCS-Mamba (Ours)	**135.80**	**5.58**	**10.47%**	**0.28**	**0.35**	**21.73%**

4.3 Qualitative Analysis and Visualization

We visualize SO_2 predictions on the 2024 test set in Fig. 4. This pollutant is strongly influenced by irregular industrial activities and is therefore difficult to forecast. DCS-Mamba captures the high-concentration cluster near the center-left region, while Time-LLM and PatchTST produce over-smoothed or blocky outputs and SimVP blurs the plume boundary.

Fig. 5 further displays absolute error maps. Earthformer, PredFormer, MMVP, and ClimaX contain large high-error patches, and SimVP still leaves localized residual clusters. DCS-Mamba presents the cleanest error map with predominantly low-error regions, visually matching the quantitative results in Table 3.

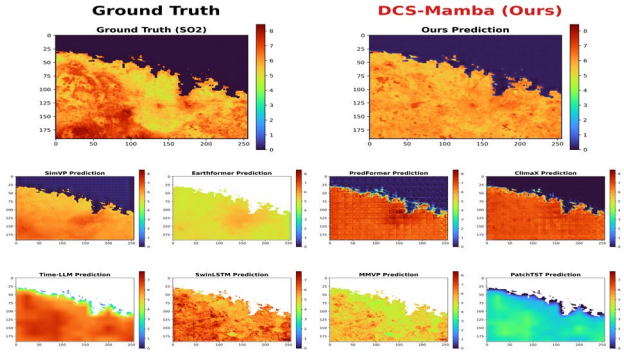


Fig. 4. SO2 forecasting visualization. DCS-Mamba recovers sharper industrial hotspots than the baseline models.

Table 4. Ablation study on the 2024 test set. Best results are bolded and colored in red; second best is colored in blue.

Model Variant	Physical Metrics			Normalized Metrics		
	MSE	MAE	MAPE	MSE	MAE	MAPE
Full Model	**163.56**	**6.29**	**18.77%**	**0.27**	**0.37**	**21.57%**
w/o U-Net (Flat)	173.66	6.62	20.71%	0.30	0.39	23.78%
w/o Text Guidance	184.30	7.10	23.65%	0.34	0.43	27.20%
w/o Mamba (Pure)	235.37	7.81	23.71%	0.53	0.52	24.74%
w/o RevIN	480.61	13.64	58.24%	2.21	1.33	75.13%

4.4 Ablation Study

To validate each component, we compare the full model with four variants: a flat Mamba stack, removal of text guidance, replacement of Mamba with 3D convolutions, and removal of RevIN. Table 4 reports the results.

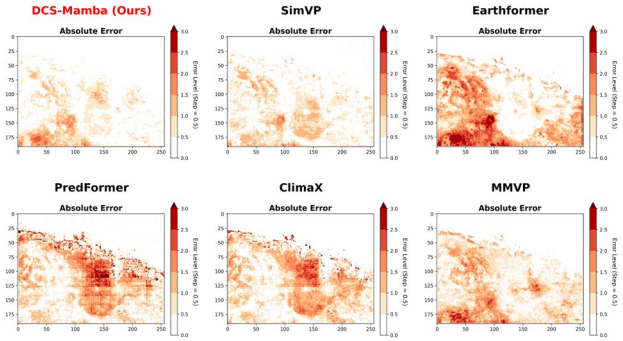


Fig. 5. Qualitative comparison of absolute error maps across seven models. Lighter regions indicate lower prediction error; DCS-Mamba shows the cleanest spatial error pattern.

The flat variant without the U-shaped hierarchy incurs a moderate performance drop, with normalized MSE increasing from 0.27 to 0.30. This suggests that multi-scale aggregation is beneficial for resolving pollution patterns at varying spatial resolutions. However, its superiority over the pure-CNN replacement highlights the intrinsic robustness of the Mamba operator itself.

Removing Text-Guided AdaGN results in clear degradation, with normalized MAPE rising to 27.20%. This confirms that meteorological prompts serve as useful external priors for events such as typhoon dispersion or winter stagnation, which are difficult to infer from numerical histories alone. The result also validates the design choice of injecting language features through normalization rather than using late fusion.

Replacing the Mamba encoder with standard convolutions causes a large physical MSE increase from 163.56 to 235.37. This validates that local convolutional receptive fields are insufficient for 45-day atmospheric dynamics, whereas selective scanning propagates long-range temporal information efficiently. The most severe degradation occurs without RevIN, where normalized MSE rises to 2.21 and MAPE reaches 75.13%. This empirically confirms the strong non-stationarity of air-quality observations and the necessity of reversible distribution alignment before physical-scale reconstruction.

5 Conclusion and Future Work

We presented DCS-Mamba, a framework for efficient, high-resolution spatiotemporal pollutant forecasting. By combining Mamba-based selective scanning, trend-seasonal RevIN alignment, and Text-Guided AdaGN, the model captures long-term pollutant dynamics while incorporating meteorological semantics. Extensive experiments on large-scale real-world datasets show that DCS-Mamba outperforms competitive video-prediction, Transformer, and foundation-model baselines, while preserving fine-grained spatial details. The results also show stable generalization across coastal industrial and mountainous terrain regimes, indicating that the framework is not tied to a single local topography. Future work will integrate real-time meteorological streams,

scale the framework to broader geographical regions, and add uncertainty quantification for extreme pollution events.

References

1. Landrigan, P.J.: Air pollution and health. Lancet Public Health 2(1), e4–e5 (2017)
2. Zeng, X., Pielke, R.A., Eykholt, R.: Chaos theory and its applications to the atmosphere. Bull. Am. Meteorol. Soc. 74(4), 631–644 (1993)
3. Muthukumar, P., et al.: PM2.5 air pollution prediction through deep learning using multi-source meteorological, wildfire, and heat data. Atmosphere 13(5), 822 (2022)
4. Kalaiselvi, S., Anitha, V., Manimaran, V., Lawrence, T.S.: Air quality prediction using multi-source remote sensing data integration with hybrid deep learning framework. Scientific Reports (2025)
5. Shi, X., et al.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. NeurIPS 28 (2015)
6. Gao, Z., Tan, C., Wu, L., Li, S.Z.: SimVP: Simpler yet better video prediction. In: CVPR, pp. 3170–3180 (2022)
7. Zhong, Y., Liang, L., Zharkov, I., Neumann, U.: MMVP: Motion-matrix-based video prediction. In: ICCV, pp. 4320–4330 (2023)
8. Gao, Z., et al.: Earthformer: Exploring space-time transformers for earth system forecasting. NeurIPS 35, 25390–25403 (2022)
9. Tang, Y., Qi, L., Li, X., Ma, C., Yang, M.H.: Video prediction transformers without recurrence or convolution. TMLR (2026)
10. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: FlashAttention: Fast and memory-efficient exact attention with IO-awareness. NeurIPS 35, 16344–16359 (2022)
11. Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J.K., Grover, A.: ClimaX: A foundation model for weather and climate. arXiv:2301.10343 (2023)
12. Jin, M., et al.: Time-LLM: Time series forecasting by reprogramming large language models. In: ICLR (2024)
13. Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: Long-term forecasting with transformers. In: ICLR (2023)
14. Hameed, S., et al.: Deep learning based multimodal urban air quality prediction and traffic analytics. Scientific Reports 13(1), 22181 (2023)
15. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First Conference on Language Modeling (2024)
16. Zhu, L., et al.: Vision Mamba: Efficient visual representation learning with bidirectional state space model. arXiv:2401.09417 (2024)
17. Kim, T., et al.: Reversible instance normalization for accurate time-series forecasting against distribution shift. In: ICLR (2022)
18. Wu, H., Xu, J., Wang, J., Long, M.: Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. NeurIPS 34, 22419–22430 (2021)
19. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 (2019)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI, pp. 234–241 (2015)