Sensitivity of Transformer-Based $PM_{2.5}$ Forecasting to Input Sequence Length

Author Information

- Kwon Jang
- rnjsdl0127@gmail.com
- 201-1, Suri Building, 22 Samdeok-ro 37beon-gil, Manan-gu, Anyang-si, Gyeonggi-do, Republic of Korea

Kwon Jang[1], Seung-Hee Han[1], Kyung-Hui Wang[1], and Hui-Young Yun[2]

[1]Department of Environmental Engineering, Anyang University, Anyang, Gyeonggi, Republic of Korea
[2]Department of Environmental and Energy Engineering, Anyang University, Anyang, Gyeonggi, Republic of Korea

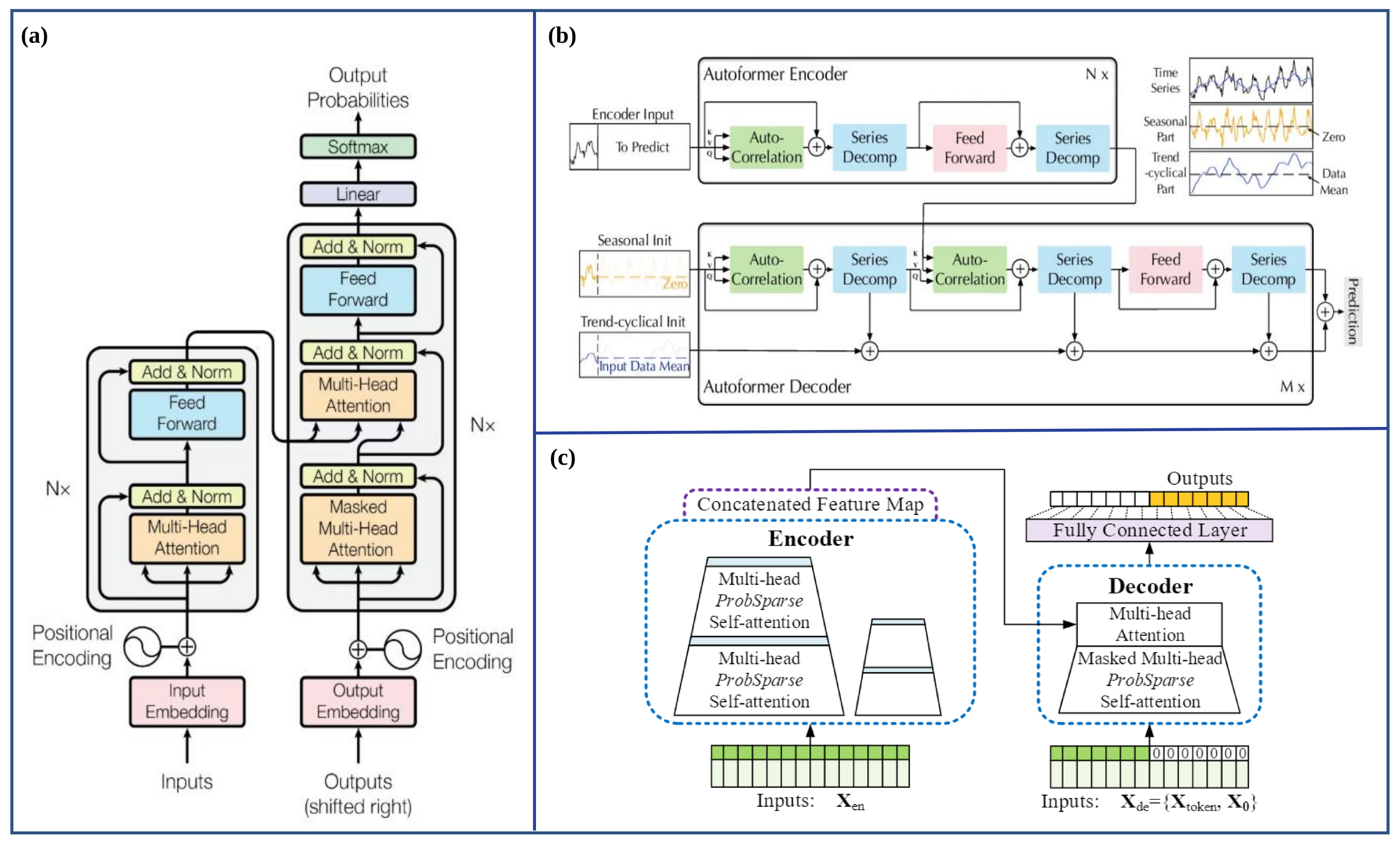
ANYANG UNIVERSITY 1948

Sharing is encouraged

Introduction

- ❖ Accurate $PM_{2.5}$ prediction is essential for air quality management and high-concentration event mitigation, particularly in Seoul where high emissions, complex urban structure, seasonal meteorology, and long-range transport increase prediction uncertainty.
- ❖ Transformer-based models are widely used for air pollution forecasting; however, the effect of input sequence length has not been sufficiently investigated.
- ❖ In this study, three Transformer architectures—Vanilla Transformer, Autoformer, and Informer—were compared under identical conditions, and the impact of input sequence length (72h, 168h, 360h) on short-term (24h) and long-term (72h) $PM_{2.5}$ prediction performance was analyzed.

Model Overview



< Fig 1. Architecture of three Transformer-based models used in this study:
(a) Vanilla Transformer (Vaswani et al., 2017), (b) Autoformer (Wu et al., 2021), (c) Informer (Zhou et al., 2021)>

(a) Vanilla Transformer - VT (Vaswani et al., 2017)

- Uses the full self-attention mechanism to learn dependencies across all time steps.
- Computational complexity grows quadratically with input length, i.e., $O(L^2)$.
- Used as the baseline model in this study.

(b) Autoformer - AF (Wu et al., 2021)

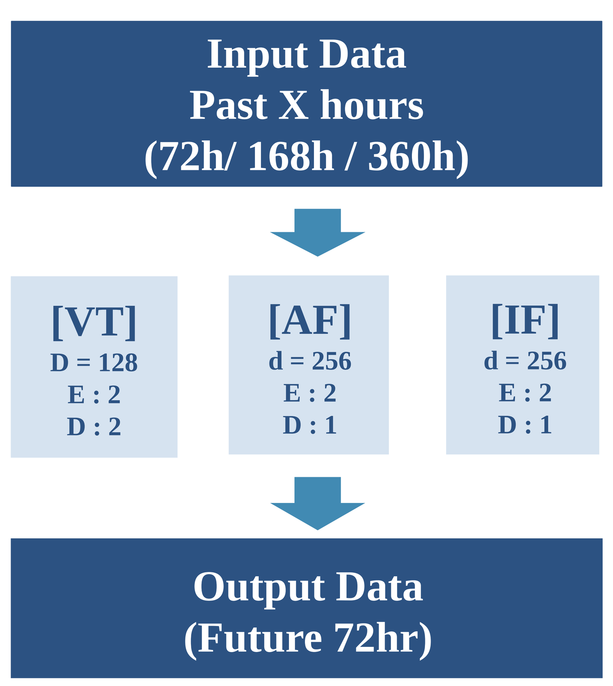
- Replaces conventional attention with Auto-Correlation and decomposes the time series into trend and seasonal components, enabling effective learning of long-term patterns.

(c) Informer - IF (Zhou et al., 2021)

- Employs ProbSparse Self-Attention to reduce complexity to $O(L \log L)$.
- Compresses the encoder sequence through a Distilling operation.
- Designed for long-sequence forecasting.

→ Applying identical input lengths allows the effects of model structure and input length to be analyzed separately.

Methodology



<Fig 2. Workflow Diagram>

- Three Transformer models (VT, Autoformer, Informer) were evaluated under three input sequence lengths (3-day, 7-day, 15-day) and two forecast horizons (24h, 72h).
- Each model's architecture was hyperparameters were individually optimized through tuning.
- All other settings (training data, preprocessing and learning strategy) were fixed identically to isolate the effect of input sequence length.

Study Area & Data

- 25 monitoring stations in Seoul
- 2018 - 2024 Hourly time series data
- Train : 2018 - 2022
- Validation : 2023
- Test : 2024

<Table 1. Hyperparameter settings used in model training>

	VT	Autoformer	Informer
D_model	128	256	256
N_heads	8	4	4
Encoder	2	2	2
Decoder	2	1	1
D_ff	256	512	512
Dropout	0.3	0.3	0.3
Batch size	32	64	32
Learning Rate	1e-5	1e-5	1e-5

<Table 2. Types of input arguments used>

Meteorological Factors		
Temperature	Precipitation	Wind Speed
Wind Direction	Relative Humidity	Atmospheric Pressure
Air Quality Data		
NO_2, SO_2, O_3, CO, $PM_{2.5}$, PM_{10}		
Temporal Encoding		
Hour(sin/cos), Day of week(sin/cos), Day of Year(sin/cos)		

Settings

- Scaling : Min-Max
- Missing : Precipitation → 0 mm / Others → MICE
- Optimizer : Adam
- Loss : MSE Loss
- Early Stopping : Yes(max 1000 epoch)

<Table 3. Actual Epochs (Early Stopping)>

	3-day	7-day	15-day
VT	48	18	51
Autoformer	39	36	38
Informer	40	35	55

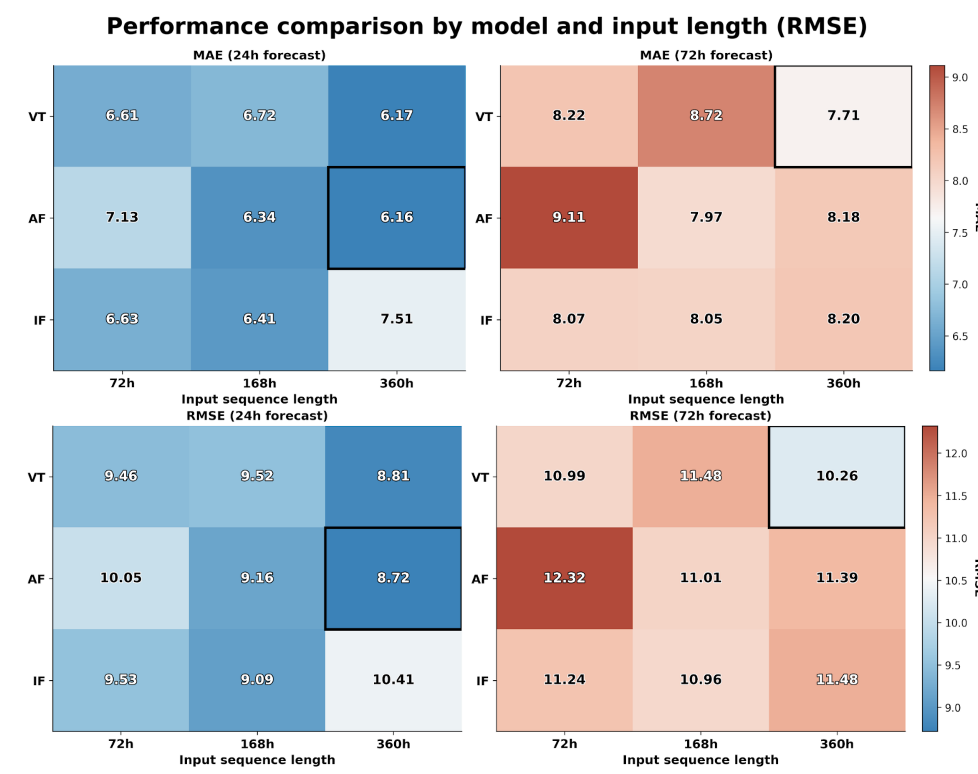
Performance Metrics

<Table 4. Description of evaluation metrics for model performance>

MAE	RMSE	R	POD	FAR
Mean Absolute Error	Root mean squared error	Correlation Coefficient	Probability of Detection	False Alarm Rate

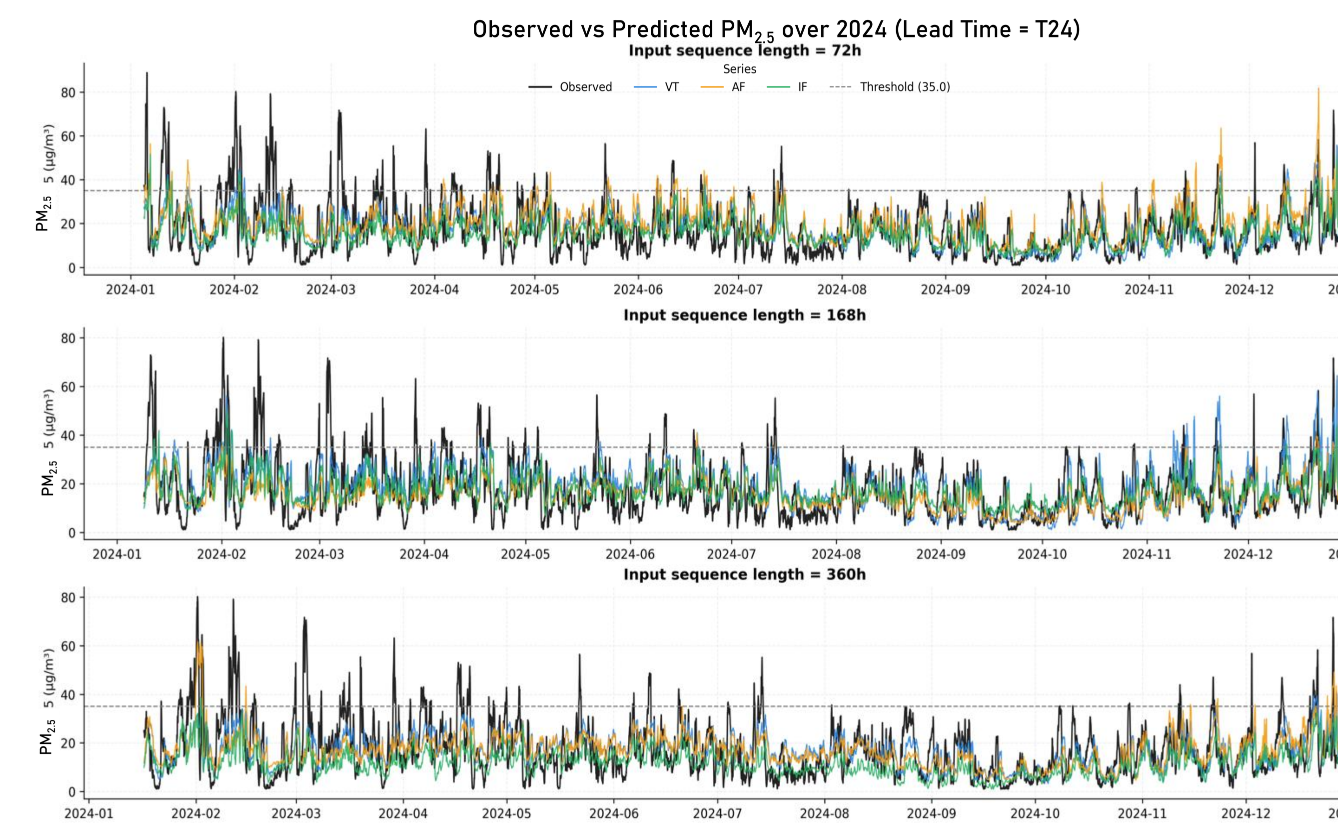
Results

Overall Performance



< Fig 3. Overall Performance>

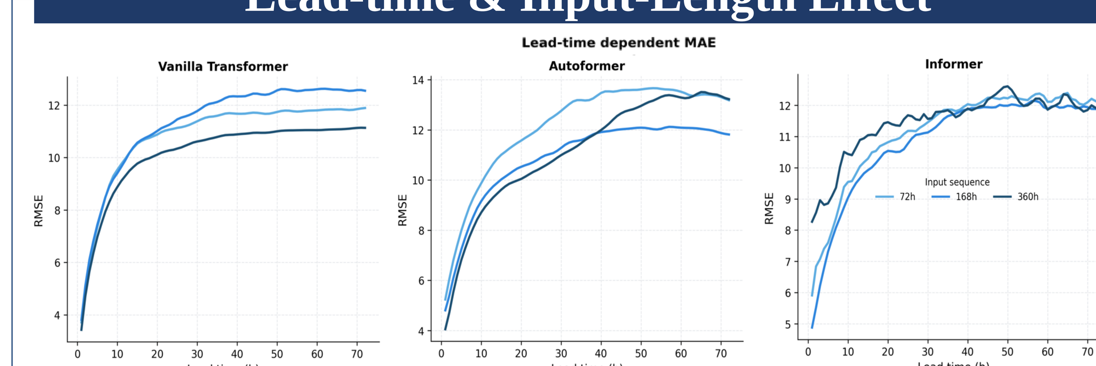
- In the 24-hour forecast, AF-360h achieved the lowest errors (MAE 6.16, RMSE 8.72), followed closely by VT-360h (MAE 6.17, RMSE 8.81).
- In the 72-hour forecast, VT-360h showed the best overall performance (MAE 7.71, RMSE 10.26).
- For both AF and IF, the 168-hour input consistently yielded lower errors than 360h in the 72-hour forecast.
- Notably, the performance gap between 168h and 360h was larger for AF (RMSE difference: 0.38) than for IF (0.52), suggesting that sensitivity to input length varies across model architectures.
- These results indicate that increasing input length does not necessarily improve long-term prediction performance.



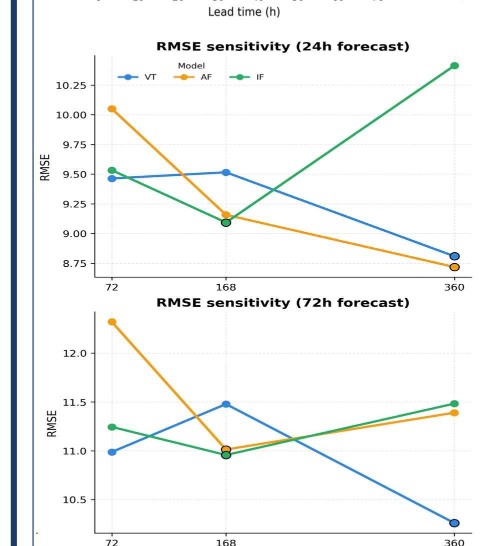
< Fig 4. Annual Time-Series Comparison (T24)>

- All models captured the overall seasonal variability of $PM_{2.5}$ throughout 2024.
- The 360-hour input produced the most stable predictions but tended to underestimate sharp concentration peaks, while the 72-hour input better reproduced peaks but with greater variability.
- This indicates a trade-off between prediction stability and peak reproducibility depending on input length.

Lead-time & Input-Length Effect



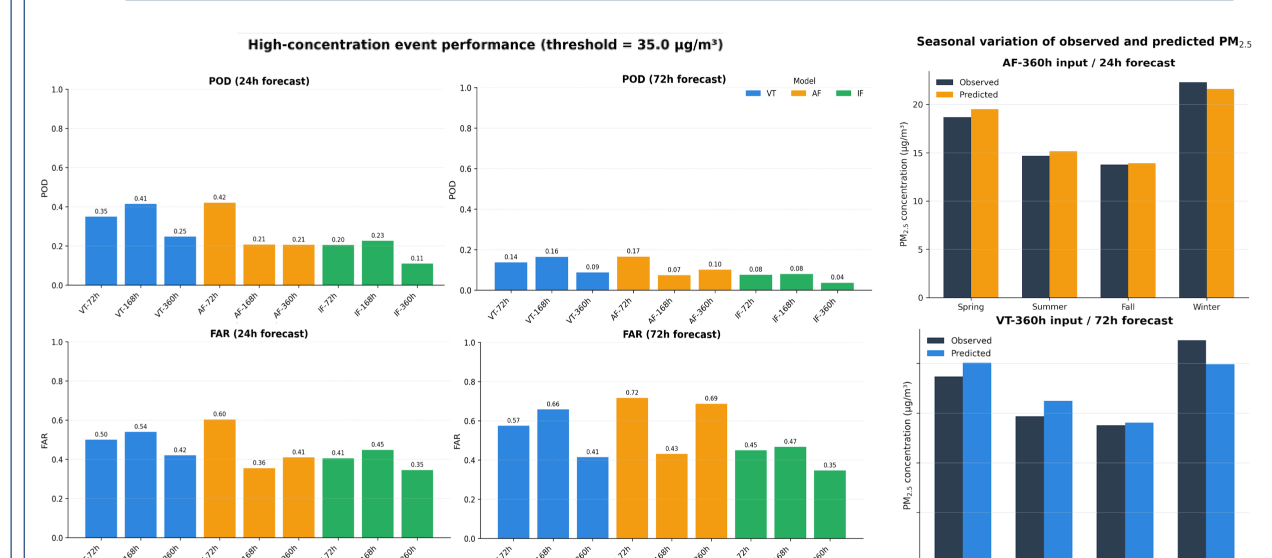
< Fig 5. Lead-Time Effect>



< Fig 6. Input-Length Sensitivity>

- Errors increased with longer lead times, but the effect of input length varied by model.
- VT: Best performance at 360h for both forecasts (24h RMSE: 8.81, 72h: 10.26), with a non-monotonic pattern.
- AF: Improved up to 168h, but 360h improved only 24h forecast (RMSE 8.72) and worsened 72h (RMSE 11.39).
- IF: Optimal at 168h (24h RMSE: 9.09, 72h: 10.96); 360h increased errors for both.

Seasonal & High-Pollution Events



< Fig 7. Seasonal & High-Pollution Events>

- South Korea showed clear seasonal $PM_{2.5}$ patterns, with spring overestimation and winter underestimation.
- A POD–FAR trade-off was observed for high-$PM_{2.5}$ events (≥35 µg/m³).
- AF-72h detected more events but produced more false alarms.
- AF-360h and VT-360h captured seasonal patterns, but winter peaks were underestimated.

Conclusion

- ❖ This study demonstrates that input sequence length is a key factor influencing the performance of Transformer-based $PM_{2.5}$ prediction.
- ❖ For short-term (24h) prediction, AF-360h was the most suitable as it stably captured trend and seasonal patterns.
- ❖ For long-term (72h) prediction, VT-360h showed the best performance by effectively learning global temporal dependencies.
- ❖ Therefore, input sequence length should be selected according to model structure and forecasting objective, not simply "the longer, the better."

Acknowledgements

This research was supported by Particulate Matter Management Specialized Graduate Program through the Korea Environmental Industry & Technology Institute(KEITI) funded by the Ministry of Climate, Energy and Environment(MCEE)