# Supplementary Material for Swin-AQI

Van Truong Thinh Nguyen[1,2], Thanh-Hai Tran[1,2], and Xuan-Bach Le[1,2⋆]

[1] Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
[2] Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam
{thinh.nguyen0334,tthai.sdh241,lexuanbach}@hcmut.edu.vn

## 1 Extended Method Details

This section expands the physics-consistency term $\mathcal{L}_{\text{physics}} = \mathcal{L}_{\text{BL}} + \mathcal{L}_{\text{K}}$ from the main paper's Training Objective. Both sub-terms are weak image-proxy regularizers, not calibrated radiometric supervision: they never replace the SILAM labels, only nudge predictions to stay consistent with the observed image brightness.

The spectral extinction coefficient $\beta = \alpha_{\text{fine}} \cdot \hat{c}_{\text{PM}_{2.5}} + \alpha_{\text{coarse}} \cdot m_{\text{coarse}} + \beta_{\text{Rayleigh}}$ follows the particulate structure of the modified IMPROVE equation [1]: the fine-mode term scales with the predicted $\text{PM}_{2.5}$ concentration directly, the coarse-mode term scales with the coarse mass implied by $\text{PM}_{10} - \text{PM}_{2.5}$, and $\beta_{\text{Rayleigh}}$ is a constant floor for molecular (non-particulate) scattering. All three constants ($\alpha_{\text{fine}} = 3.0$, $\alpha_{\text{coarse}} = 0.6$, $\beta_{\text{Rayleigh}} = 10.0$) are fixed, not learned, so the model cannot inflate or deflate the physics term by adjusting them.

The Beer–Lambert branch converts this extinction coefficient into a predicted transmittance $\hat{T} = \exp(-\beta \times 10^{-3} \times d_{\text{km}})$ over the scene-relative path length $d_{\text{km}}$ derived from MiDaS inverse depth (Eq. 7 in the main paper), and matches it against $\tilde{V}$, the spatial mean of the processed image tensor the backbone consumes, via $\mathcal{L}_{\text{BL}} = \|\hat{T} - \tilde{V}\|_1$. The Koschmieder branch instead derives a visibility estimate $\hat{V} = \text{clip}(K/\beta \times 10^3/100,\ 0,\ 1)$ from the same extinction coefficient, with $K = 3.0$ a fixed contrast-threshold constant (so that the resulting visibility $K/\beta$ is converted from Mm to km via $\times 10^3$, then normalized by the 100 km clear-air reference), and matches it against the same brightness proxy via $\mathcal{L}_{\text{K}} = \|\hat{V} - \tilde{V}\|_1$. Using one shared brightness target $\tilde{V}$ for both terms means the two branches agree or disagree on the same evidence, rather than each pulling the prediction toward a separately-defined notion of visibility.

Because $d_{\text{assumed}} = 1\,\text{km}$ is a fixed scene-scale hyperparameter rather than a calibrated distance, and $\tilde{V}$ is derived from the processed image tensor rather than a photometrically calibrated luminance, both loss terms are best read as weak consistency checks: they penalize particulate predictions that are grossly incompatible with how dark or washed-out the image already looks, without claiming to recover physical visibility in kilometers.

⋆ Corresponding author

## 2 Algorithm

Algorithm 1 gives the main paper's Figure 1 pipeline as pseudocode.

**Algorithm 1** Swin-AQI training and inference

**Require:** RGB image $\mathbf{I}$, capture hour $t$, pollutant labels $\mathbf{y}$ during training
**Ensure:** Estimates $\hat{\mathbf{y}} = [\hat{y}_{\mathrm{PM}_{2.5}}, \hat{y}_{\mathrm{PM}_{10}}, \hat{y}_{\mathrm{CO}}, \hat{y}_{\mathrm{NO}_2}, \hat{y}_{\mathrm{SO}_2}, \hat{y}_{\mathrm{O}_3}]$

1: Estimate MiDaS relative depth $D$ and pool it to the post-backbone spatial resolution
2: Tokenize $\mathbf{I}$ and process tokens with shifted-window Swin-T blocks
3: Apply $B_{\mathrm{depth}}$ within the post-backbone depth-aware attention module
4: Encode capture hour as Fourier features $\phi(t)$ and concatenate with pooled visual features
5: Predict six pollutant outputs with pollutant-specific regression heads
6: During training, minimize uncertainty-weighted smooth L1 loss plus PM-ordering and weak visibility-consistency terms
7: **return** $\hat{\mathbf{y}}$

## 3 Label Degeneracy

The main paper (Section 3.1) notes that images are labeled via nearest-neighbor lookup against the SILAM global operational product (v6.1, natively gridded at 20 km resolution). Because the main-paper study area has a 5 km radius, it is small relative to this native grid spacing, so images falling within the same grid cell and hour can receive identical six-pollutant label vectors. This section quantifies that degeneracy.

Table 1 reports, over the full set of 1,977 images: the number of distinct SILAM grid cells covered, the number of distinct SILAM hourly timestamps, the number of distinct (cell, hour) label combinations, and the fraction of images that share an identical label vector with at least one other image in the dataset.

**Table 1.** Label degeneracy over the study area.

| Quantity | Value |
|---|---|
| Distinct SILAM grid cells covered | 2 |
| Distinct SILAM hourly timestamps | 33 |
| Distinct (cell, hour) label combinations | 35 |
| Fraction sharing an identical label vector | 99.9% |

The study area spans two SILAM grid cells, but the sample is heavily skewed toward one of them: 1,650 of 1,977 images (83.5%) fall in a single cell, with the

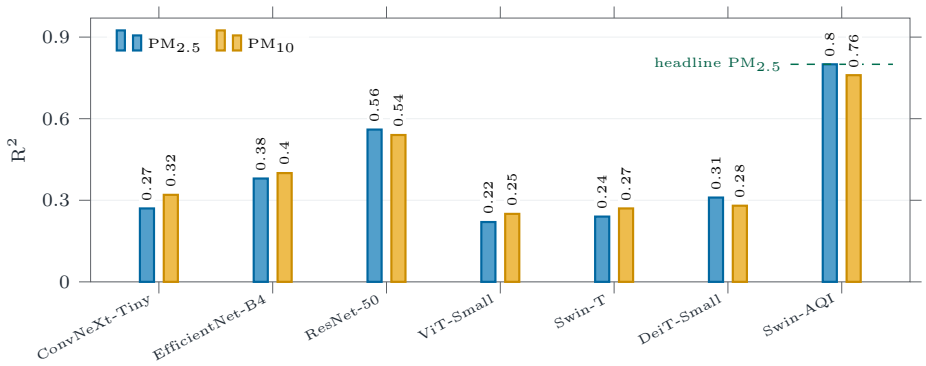


**Fig. 1.** Particulate $R^2$ comparison across baselines (main paper, Table 1).

remaining 327 (16.5%) in the second. Combined with 33 distinct SILAM hourly timestamps (spanning multiple dates over the collection window), the dataset contains only 35 distinct label combinations for 1,977 images, and 99.9% of images share their exact six-pollutant label vector with at least one other image. In practice, the label is close to a function of the capture datetime-hour alone. This bounds what the time-embedding ablation (Section 4.2 of the main paper) can be read as evidence for: the large drop from removing the time embedding is expected given this near-total spatial degeneracy.

## 4 Additional Results and Diagnostics

Figure 1 re-plots the baseline $R^2$ comparison (main paper, Table 1) as bars, and Figure 2 does the same for the ablation comparison (main paper, Table 3), so the relative gaps are easier to read at a glance than in the tabular form.

The most controlled same-family contrast in Figure 1 is between Swin-AQI and standard Swin-T ($R^2$=0.24/0.27 for $PM_{2.5}$/$PM_{10}$): the two share a backbone family, but Swin-AQI additionally receives the capture-hour feature, uncertainty-weighted multi-task head, depth-aware attention, and physics-informed training — all of which are absent from the Swin-T baseline (main paper, Section 3.2). The gap to Swin-AQI's $R^2$=0.80/0.76 is therefore attributable to the full Swin-AQI recipe; Table 3 of the main paper gives the per-component breakdown, with time embedding and uncertainty weighting accounting for the largest individual shares. ResNet-50 is the closest baseline ($R^2$=0.56/0.54), while ViT-Small trails every other model ($R^2$=0.22/0.25).

Figure 2 shows the same drops the main paper discusses: removing the time embedding costs the most ($R^2$ falls from the 0.79 full-reference to 0.58, a 0.21 drop), uncertainty weighting is close behind (0.60, a 0.19 drop), and depth-aware attention and the physics loss contribute smaller, separable gains (0.69 and 0.72, i.e. 0.10 and 0.07 drops). The main paper's Results section reports

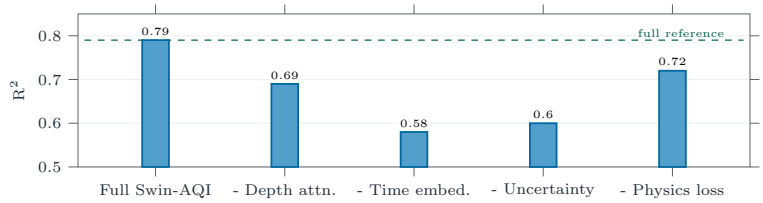


**Fig. 2.** Leave-one-out ablation summary for $PM_{2.5}$ (main paper, Table 3, shared 50-epoch budget).

the resulting $R^2$ and MAE for the primary targets, and the $R^2$ for the macro-averaged auxiliary gases.

## 5 Expanded Limitations

**Label dependence.** The image labels come from the nearest SILAM grid cell and hour (Section 3), not from direct physical measurements. These fields have not been validated against independent reference-station readings, so their point-level accuracy is unknown; if SILAM misses a street-scale source, the image model can inherit that bias. Station or mobile-sensor validation of the label source itself is therefore needed before regulatory or neighborhood-scale exposure claims.

**Physics-loss brightness scale.** The observed-brightness proxy $\tilde{V}$ (Section 1) is derived from the image tensor that the training pipeline processes. Because the implementation applies brightness normalization twice (dividing by 255 before the loss), $\tilde{V}$ ends up numerically much smaller than a directly normalized $[0, 1]$ raw-brightness value. Consequently, both physics terms are matched against an uncalibrated scale rather than photometric radiance, and operate as an architectural regularizer whose numerical operating point has not been independently checked, rather than as a calibrated photometric constraint.

**Depth reliability.** MiDaS provides relative, not metric, depth. Reflective glass, blank sky, dense vegetation, an unusual camera pitch, or rain on the lens can all distort the depth prior, which propagates to both the depth-aware attention and the physics-consistency terms.

**Weather and lighting.** Many everyday conditions unrelated to particulates still lower visibility: fog, mist, backlighting, hard shadows, and exposure shifts. RGB imagery alone cannot reliably separate these from genuine aerosol haze, so predictions under such conditions should be treated cautiously.

**Gas scope.** The model predicts CO, $NO_2$, $SO_2$, and $O_3$ as auxiliary outputs, but their performance is contextual rather than physical. Traffic-heavy or industrial scenes can correlate with gases, yet RGB images do not provide reliable gas absorption measurements. Operational claims should therefore remain focused on $PM_{2.5}$ and $PM_{10}$.

**Generalization and deployment.** The study is limited to one city, a two-week daylight window, and mostly moderate particulate levels, and the model leans partly on the capture hour rather than on image evidence alone. Multi-city/seasonal tests, held-out-hour controls, and fusion with calibrated sensors are the next steps; camera-based sensing should complement, not replace, reference stations.

## 6 Appendix

Source code:
https://github.com/ZleeMb0334/acivs2026_swin-aqi
Dataset:
https://www.kaggle.com/datasets/thnhnguynvntrng/swin-aqi-dataset

## References

1. Hand, J.L., Copeland, S.A., Day, D.E., Dillner, A.M., Indresand, H., Malm, W.C., McDade, C.E., Moore Jr., C.T., Pitchford, M.L., Schichtel, B.A., Watson, J.G.: Spatial and seasonal patterns and temporal variability of haze and its constituents in the United States: Report V. Tech. rep., Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO (2011)