

A USAGE OF LLMs

All authors declare that the LLMs are used as a general-purpose assist tool for polishing the manuscript, LLMs do not contribute to research ideation.

B DATASET

Shanghai Radar: The Shanghai Radar dataset (Chen et al., 2020) contains radar scans images from the WSR-88D radar in Pudong, Shanghai, China, collected between October 2015 and July 2018. Each scan covers a 501×501 km area at 6-minute intervals. The data are segmented into 25-frame sequences with a stride of 20, and are split into training, validation, and test subsets following (Chen et al., 2020). The first 5 frames (30 minutes) are used to predict the next 20 frames (120 minutes). Radar values are rescaled to $[0, 70]$, and CSI/HSS are computed at thresholds $[20, 30, 35, 40]$.

SEVIR: The SEVIR dataset (Veillette et al., 2020) consists of storm events from 2017 to 2020, collected every 5 minutes over 4-hour windows using GOES-16 and NEXRAD sources. Each event spans a 384×384 km region. We use the vertically integrated liquid (VIL) radar mosaics and extract 25-frame sequences with a stride of 12 following (Yu et al., 2024). The first 5 frames (25 minutes) are used to predict the next 20 frames (100 minutes). The dataset is split into training, validation, and test subsets using cutoff dates: before 2019-01-01 for training, 2019-01-01 to 2019-06-01 for validation, and 2019-06-01 to 2020-12-31 for testing. Radar values are rescaled to $[0, 255]$, and CSI/HSS are evaluated at thresholds $[16, 74, 133, 160, 181, 219]$.

MeteoNet: MeteoNet (Larvor et al., 2020) provides radar and auxiliary meteorological data over two regions in France for the years 2016–2018. We use rain radar scans over north-western France, available at 6-minute intervals. Sequences of 25 frames are extracted using a stride of 12, where the first 5 frames (30 minutes) are used to predict the next 20 frames (120 minutes). The data are partitioned into training, validation, and test sets with cutoff dates of 2016-01-01 to 2017-12-31, 2018-01-01 to 2018-06-01, and 2018-06-01 to 2018-12-31, respectively. Radar values are rescaled to $[0, 70]$, and CSI/HSS are computed at thresholds $[12, 18, 24, 32]$.

CIKM: The CIKM dataset² from CIKM AnalytiCup 2017 provides 15-frame radar echo sequences sampled every 6 minutes over a 1.5-hour period, covering a 101×101 km region in Guangdong, China. Each sample includes reflectivity maps at four altitudes from 0.5 km to 3.5 km; we use data at the 2.5 km level. Different from previous datasets, in CIKM, first 5 frames (30 minutes) are used to predict the next 10 frames (60 minutes). The dataset is split into training, validation, and test subsets using the official partition. Pixel values are rescaled to $[0, 70]$, and CSI/HSS metrics are computed at thresholds $[20, 30, 35, 40]$.

C TOKEN-WISE ATTENTION: TIME-SPACE COMPLEXITY

C.1 SELF-ATTENTION (ViT (DOSOVITSKIY ET AL., 2021))

Given input embeddings $z \in \mathbb{R}^{n \times d}$ (with n tokens and width d), we form $Q = zW_Q$, $K = zW_K$, $V = zW_V$, where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$. The scaled dot-product attention is:

$$\hat{z} = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \quad (16)$$

Projecting to Q, K, V costs $3O(nd^2)$. Computing logits QK^\top is $O(n^2d)$, the softmax is $O(n^2)$, and multiplying by V is $O(n^2d)$. Thus, the overall time and space complexities are:

$$T(n, d) = O(nd^2 + n^2d) \quad (17)$$

$$S(n, d) = O(nd) + O(n^2) = O(n^2 + nd) \quad (18)$$

For a feature map of size $h \times w$, $n = hw$ and $d \ll n$, equations. 17–18 simplify to:

$$T(n, d) = O(n^2), \quad S(n, d) = O(n^2). \quad (19)$$

²<https://tianchi.aliyun.com/dataset/1085>

C.2 TOKEN-WISE ATTENTION

We analyze the TWA operations through each step (excluding the outer linear projections/MLPs):

Equations 9, 11:

$$T(n, d) = O(nd) + O(n) + O(nd) = O(nd) \quad (20)$$

$$S(n, d) = O(n) + O(d) \quad (21)$$

Equation 10:

$$T(n, d) = O(nd), \quad S(n, d) = O(nd). \quad (22)$$

Equation 12:

$$T(n, d) = O(nd), \quad S(n, d) = O(nd). \quad (23)$$

The TWA scales linearly in the number of tokens:

$$T(n, d) = O(nd) \quad (24)$$

$$S(n, d) = O(nd + n + d) = O(nd) \quad (25)$$

For a feature map of size $h \times w$, $n = hw$ and $d \ll n$, equations 24–25 simplify to:

$$T(n, d) = O(n), \quad S(n, d) = O(n). \quad (26)$$

D SPATIO-TEMPORAL ENCODER

The gradient flow of \mathcal{L} from equation 8 with respect to each input frame x_j decomposes into:

$$\frac{\partial \mathcal{L}_{123}}{\partial x_j} = \gamma \left[\frac{\partial \mathcal{L}_{23}}{\partial h_T} \left(\frac{\partial h_T}{\partial x_j} + \frac{\partial h_T}{\partial \mu} \frac{\partial \mu}{\partial x_j} \right) + \sum_{m,t} \frac{\partial \mathcal{L}_{23}}{\partial s_m^t} \frac{\partial s_m^t}{\partial x_j} \right] + (1 - \gamma) \frac{\partial \mathcal{L}_1}{\partial \mu} \frac{\partial \mu}{\partial x_j} \quad (27)$$

$$\text{with } \frac{\partial s_m^t}{\partial x_j} = \frac{\partial s_m^t}{\partial s_m^0} \frac{\partial s_m^0}{\partial \mu} \frac{\partial \mu}{\partial x_j} = -\sqrt{\alpha_t} \frac{\partial \mu}{\partial x_j}, \quad (28)$$

$$\Rightarrow \frac{\partial \mathcal{L}_{123}}{\partial x_j} = \gamma \frac{\partial \mathcal{L}_{\theta_2}}{\partial h_T} \frac{\partial h_T}{\partial x_j} + \left[\gamma \left(\frac{\partial \mathcal{L}_{23}}{\partial h_T} \frac{\partial h_T}{\partial \mu} - \sum_{m,t} \sqrt{\alpha_t} \frac{\partial \mathcal{L}_{23}}{\partial s_m^t} \right) + (1 - \gamma) \frac{\partial \mathcal{L}_1}{\partial \mu} \right] \frac{\partial \mu}{\partial x_j}, \quad (29)$$

$$\text{where } \frac{\partial h_T}{\partial x_j} = \underbrace{\left(\prod_{i=j}^{T-1} \frac{\partial h_{i+1}}{\partial h_i} \right)}_{J_{j \rightarrow T}} \frac{\partial h_j}{\partial x_j}, \frac{\partial h_T}{\partial \mu} = \underbrace{\left(\prod_{i=1}^{T-1} \frac{\partial h_{i+1}}{\partial h_i} \right)}_{J_{1 \rightarrow T}} \frac{\partial h_1}{\partial \mu}, T = T_{\text{in}} + T_{\text{out}}, j \in [1, T_{\text{in}}]. \quad (30)$$

where $\mathcal{L}_{123}, \mathcal{L}_{23}, \mathcal{L}_1$ denote $\mathcal{L}(\theta_1, \theta_2, \theta_3), \mathcal{L}(\theta_2, \theta_3), \mathcal{L}(\theta_1)$ respectively.

E VISUALIZATION

Here are additional qualitative examples across datasets. As shown in Figures 5, 6, 7, 8, all deterministic backbones become noticeably blurry by the 60-minute horizon, with high-reflectivity cores and fine-scale details fading. In contrast, RainDiff preserves sharper echoes and yields more accurate precipitation intensity and localization, especially at longer lead timesteps. Compared to DiffCast, the closest baseline, RainDiff produces less stochastic, more coherent precipitation contours while better matching the observed air masses' shape and position.

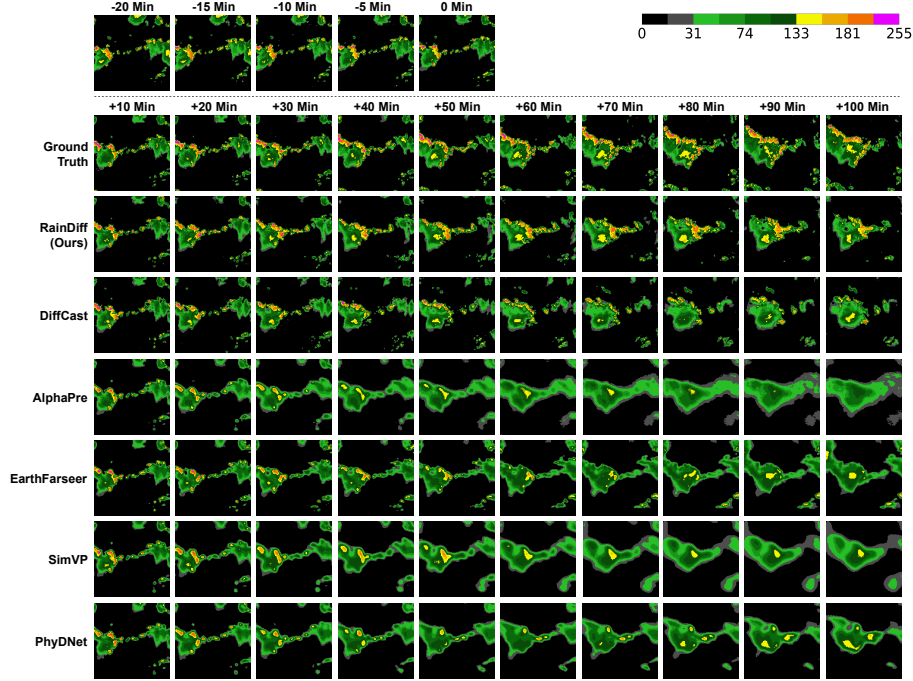


Figure 5: Prediction examples on the SEVIR dataset, where the color bar on the top right represents the reflectivity range of radar echoes.

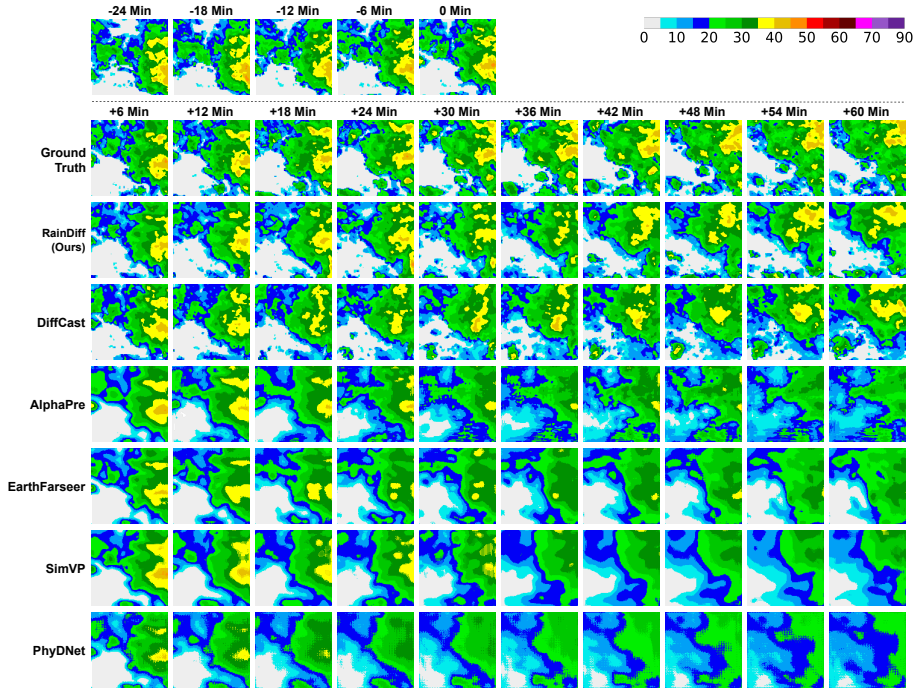


Figure 6: Prediction examples on the CIKM dataset, where the color bar on the top right represents the reflectivity range of radar echoes.

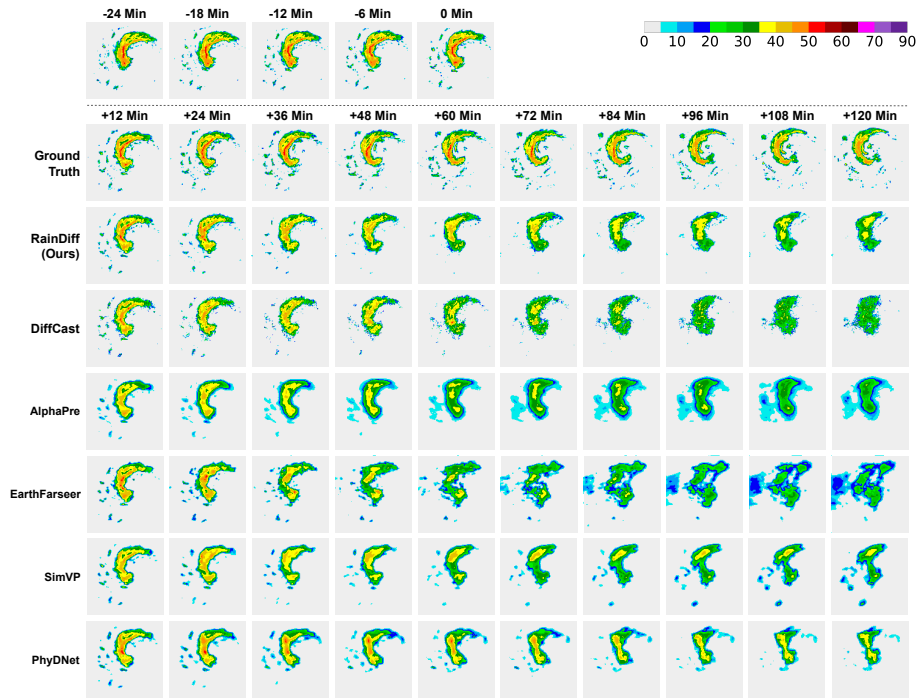


Figure 7: Prediction examples on the Shanghai Radar dataset, where the color bar on the top right represents the reflectivity range of radar echoes.

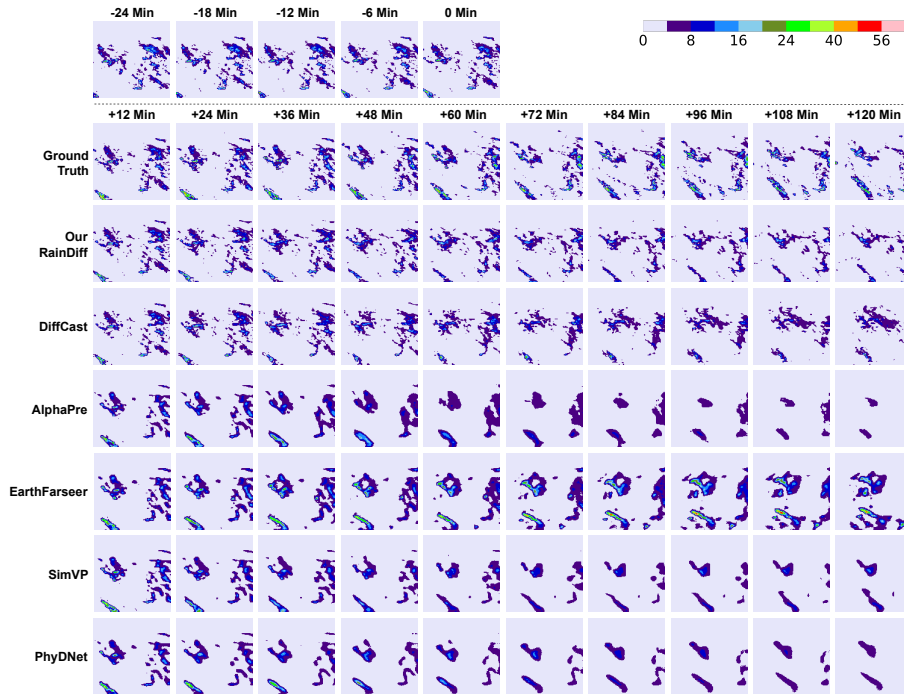


Figure 8: Prediction examples on the MeteoNet dataset, where the color bar on the top right represents the reflectivity range of radar echoes.