LEARNING FROM MULTI-PERSPECTIVES: A MULTI-MODAL PRECIPITATION NOWCASTING FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Precipitation nowcasting is a crucial spatio-temporal forecasting task with broad applications in weather warning, transportation, and agriculture. Existing deep learning based methods typically model precipitation evolution either from a radar-native perspective that captures signal-level motion and intensity dynamics, or from a visual perspective that extracts high-level morphological structures of precipitation system. Although these two perspectives are complementary, few works have integrated them into a unified framework. Moreover, recent works have introduced textual meteorological priors but only couple them with visual representations, leaving their potential synergy with radar-native dynamics largely unexplored. In this paper, we propose PreMM, a multimodal precipitation nowcasting framework that jointly models radar-native dynamics, visual representations, and textual semantic priors. PreMM consists of a Radar Dynamics Learner (RDL) that explicitly decouples radar evolution into motion and intensity components, and enhances them with frequency-domain and textual priors; a Visual Perception Learner (VPL) that reconstructs radar observations into visual representations with frequency and periodicity cues, and leverages a pre-trained vision-language encoder to capture precipitation structures; a Text Prior Generator (TPG) that produces meteorological semantic priors; and a Multimodal Fusion Network (MFN) that integrates these complementary representations for future radar prediction. Extensive experiments on multiple precipitation nowcasting benchmarks demonstrate that PreMM consistently outperforms state-of-the-art methods. Our code is available at `https://anonymous.4open.science/r/PreMM`.

1 INTRODUCTION

Precipitation nowcasting aims to predict the short-term evolution of rainfall and is crucial for severe weather warning (Pan et al., 2026; Ravuri et al., 2021), transportation scheduling (Zeng et al., 2024), and agricultural planning (Sharafkhane et al., 2024). It can be formulated as a spatio-temporal forecasting task, as radar echoes evolve across both spatial and temporal dimensions. However, unlike conventional spatio-temporal prediction tasks with relatively regular and stable dynamics (Gao et al., 2022b; Rühling Cachay et al., 2023), precipitation nowcasting involves more complex evolution characterized by irregular motion patterns, dynamic intensity variations, and complex frequency and periodic dynamics.

Existing deep learning based approaches to precipitation nowcasting can be broadly categorized into two families according to their modeling perspective. The first adopts a radar-native perspective, directly exploiting intrinsic characteristics of radar observations to characterize precipitation evolution (Li et al., 2023; Lin et al., 2025; Wu et al., 2026). For example, AlphaPre (Lin et al., 2025) disentangles radar echoes into amplitude and phase components in the frequency domain so as to model intensity and positional variations separately, whereas MoCast (Wu et al., 2026) incorporates physical knowledge of turbulent motions and decomposes them into interpretable motion components for precipitation forecasting. However, their insufficient prior guidance leaves motion and intensity modeling vulnerable to noise and irregular patterns, while simply fusing the two components at the final stage fails to fully capture their intrinsic coupling. The second adopts a visual perspective, treating radar echoes as image sequences and leveraging representation mechanisms developed in computer vision (Gao et al., 2022c; Li et al., 2024; Wang et al., 2025). Earthformer (Gao et al., 2022c) partitions radar sequences into space-time cuboids and employs attention mechanisms

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to capture their spatial and temporal dependencies, while the CNN-Transformer framework (Wang et al., 2025) combines convolutional representations with Transformer-based modeling to capture visual patterns and spatio-temporal evolution. Nevertheless, these methods overlook diverse image signal for multi-view visual modeling and employ relative small visual encoder, limiting their capacity to extract complex visual features.

These two perspectives offer complementary lenses for understanding precipitation evolution. The radar-native perspective preserves domain-specific signal characteristics that are tightly coupled with underlying physical dynamics, allowing models to capture fine-grained intensity variations and motion patterns directly from observational backscatter. In contrast, the visual perspective encodes higher-level morphological and organizational properties of precipitation systems, facilitating the recognition of echo structures and their spatio-temporal evolution. Together, these views provide a more complete characterization of precipitation processes—one rooted in raw signal continuity, the other in structured perceptual patterns. Recently, LangPrecip (Ling et al., 2026) introduces a third semantic dimension by incorporating meteorological text as a prior to constrain precipitation motion. However, its textual guidance is integrated solely with visual features, leaving its potential synergy with radar-native dynamics largely unexamined. This gap underscores the need for a unified framework that jointly models radar-native signals, visual representations, and textual semantic priors, enabling a truly holistic approach to precipitation nowcasting.

In this paper, we propose PreMM, a multimodal precipitation nowcasting framework that unifies radar-native dynamics and visual representations, guided by textual priors, to model the complex precipitation evolution. Specifically, our framework introduces a Radar Dynamics Learner (RDL) to explicitly decouple radar evolution into motion and intensity dynamics, where frequency-domain and textual priors are leveraged to enhance their respective representations, and a dual cross-attention mechanism is employed to capture their interactions. Meanwhile, a Visual Perception Learner (VPL) reconstructs radar observations into visual representations by incorporating frequency and periodicity cues, and then adopts a pre-trained vision-language model (VLM) to capture high-level precipitation structures and spatio-temporal evolution patterns beyond raw radar signals. Furthermore, we introduce a Text Prior Generator (TPG) that provides meteorological semantic guidance to complement both radar and visual representations. Finally, the Multimodal Fusion Network (MFN) integrates these complementary representations into a unified feature representation for future radar prediction. Extensive experiments across multiple precipitation nowcasting datasets demonstrate that the proposed framework consistently outperforms existing state-of-the-art methods while exhibiting strong scalability with increasing model capacity.

Our main contributions are summarized as follows:

- We propose PreMM, a multimodal precipitation nowcasting framework that unifies radar-native dynamics and visual representations with textual priors, to capture the complex evolution of precipitation system, while exhibiting favorable scalability with consistent performance gains as model capacity increases.
- We explicitly model precipitation evolution from radar-native and visual perspectives, capturing complementary dynamics rooted in raw signal continuity and structured perceptual patterns to provide a more comprehensive characterization of precipitation processes.
- We design two novel learners. RDL decouples radar evolution into motion and intensity dynamics, enriched with specific priors, and employs a dual cross-attention to model their interactions. VPL reconstructs radar observations into visual representations, and uses a pre-trained VLM encoder to capture precipitation structures beyond raw signals.

2 Related Work

2.1 Precipitation Nowcasting Models

Recent advances in deep learning have significantly improved precipitation nowcasting. Models like ConvLSTM and its variants (Shi et al., 2015; 2017b) enhance forecasting by jointly modeling temporal dependencies and spatial structures in radar echo sequences. Subsequent methods (Wang et al., 2019; Zhang et al., 2022; Li et al., 2023; 2024) further improve motion modeling, long-range dependency capturing, and multiscale representation learning, surpassing traditional extrapolation-based

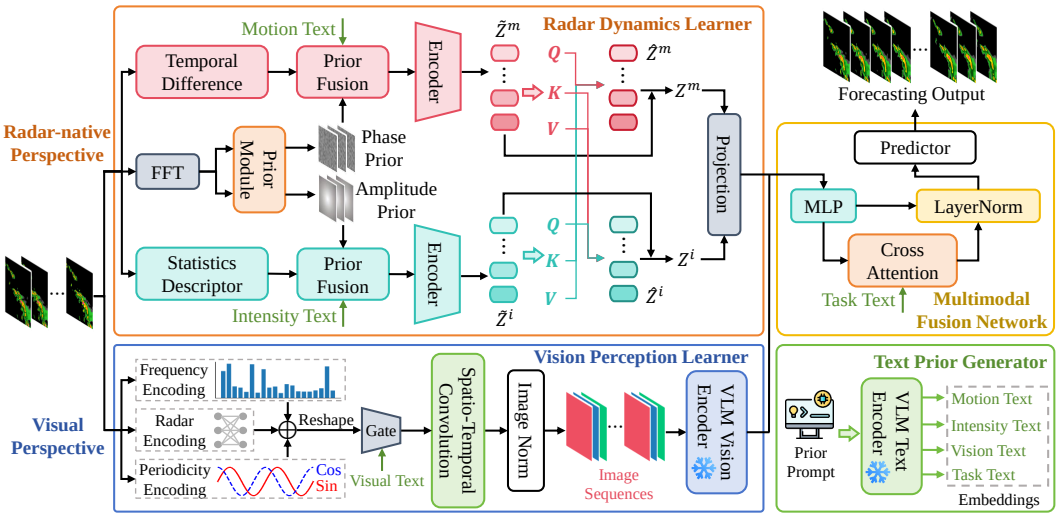

Figure 1: Overview of the PreMM architecture, which unifies radar-native dynamics and visual representations, guided by textual priors for precipitation nowcasting. Radar Dynamics Learner (RDL) disentangles radar evolution into motion and intensity dynamics, enhances them with frequency-domain and textual priors, and employs a dual cross-attention mechanism to capture their interactions. Visual Perception Learner (VPL) reconstructs radar observations into visual representations with frequency and periodicity cues, and leverages a pre-trained VLM visual encoder to capture morphological properties. Text Prior Generator (TPG) provides meteorological semantic priors to complement both radar and visual representations. Finally, Multimodal Fusion Network (MFN) integrates these complementary representations for future radar prediction.

approaches. Recent studies have adopted probabilistic and generative forecasting frameworks (Yu et al., 2024; Gong et al., 2025; Perrone Ribeiro & Faganeli Pucer, 2026) to better handle uncertainty and ambiguity in future precipitation evolution, yielding sharper, more realistic forecasts, particularly for heavy rainfall and longer prediction horizons.

2.2 Multimodal Modeling for Meteorological Patterns

Multimodal learning has emerged as a promising direction for modeling meteorological patterns. WeatherQA (Ma et al., 2024) combines diverse meteorological images with expert-written textual analyses to evaluate multimodal reasoning about severe weather development and affected areas. CLLMate (Li et al., 2025) aligns meteorological raster data with textual event descriptions to support weather and climate event forecasting. ClimateIQA (Chen et al., 2025) pairs meteorological heatmaps with textual instructions to adapt VLM for meteorological anomaly analysis. Recently, LangPrecip (Ling et al., 2026) explores a language-aware multimodal paradigm for precipitation nowcasting by introducing meteorological text as semantic guidance for modeling precipitation evolution. Specifically, textual priors are incorporated to constrain precipitation motion, but are solely coupled with visual representations. Therefore, how to effectively coordinate radar-native, visual, and textual representations within a unified framework still warrants further research.

3 Methodology

3.1 Problem Definition

We denote the input radar sequence as $X \in \mathbb{R}^{L\times C\times H\times W}$, where L is the number of the historical frames, C is the number of measurements at each space-time coordinate, H and W denote the spatial resolution of each frame respectively, and X_l represents the l-th frame. Similar to the radar setting of previous works, we set $C = 1$. The objective of precipitation nowcasting is to predict the future radar sequence $Y \in \mathbb{R}^{T\times C\times H\times W}$, where T is the length of the future frames.

3.2 RADAR DYNAMICS LEARNER

Radar Dynamics Learner (RDL) models precipitation evolution from a radar-native perspective by explicitly capturing the motion and intensity dynamics. RDL disentangles radar evolution into two components through temporal and statistical operations, and enhances them with frequency-domain and textual priors. A dual cross-attention is further introduced to capture the interactions between motion and intensity representations, producing a unified representation for subsequent forecasting.

Phase and Amplitude Priors. We first derive phase and amplitude priors to provide explicit cues for motion and intensity modeling. Given the l-th radar frame X_l, we apply a two-dimensional Fast Fourier Transform (FFT) along the spatial dimensions to obtain its frequency representation:

$$\mathcal{F}_l = \mathrm{FFT}(X_l) = \mathcal{R}_l + j\mathcal{I}_l, \tag{1}$$

where $\mathcal{R}_l$ and $\mathcal{I}_l$ denote the real and imaginary components. The corresponding amplitude and phase spectra are computed as: $\mathcal{A}_l = |\mathcal{F}_l| = \sqrt{\mathcal{R}_l^2 + \mathcal{I}_l^2}, \mathcal{P}_l = \arctan(\mathcal{I}_l, \mathcal{R}_l)$. We perform this decomposition for all frames and obtain the spectral sequences: $\mathcal{A} = \{\mathcal{A}_l\}_{l=1}^{L}, \mathcal{P} = \{\mathcal{P}_l\}_{l=1}^{L}$.

The phase and amplitude spectra provide complementary cues for radar evolution. Phase is sensitive to spatial displacement and structural changes, while amplitude reflects the energy distribution and strength of radar echoes. Therefore, we employ a lightweight Prior Module to separately encode the two spectral components: $P^m = \mathcal{G}_p(\mathcal{P}), P^i = \mathcal{G}_a(\mathcal{A})$, where $\mathcal{G}_p(\cdot)$ and $\mathcal{G}_a(\cdot)$ denote the corresponding transformations in the Prior Module. The phase prior P^m and amplitude prior P^i are introduced into the motion and intensity branches, respectively, to guide radar dynamics modeling.

Motion Branch. To capture the displacement and structural evolution of precipitation systems, we first extract inter-frame variations through temporal differencing:

$$D_l^m = X_l - X_{l-1}, \qquad l = 2, \ldots, L. \tag{2}$$

The resulting motion cues are further enhanced by the phase prior P^m, which provides complementary frequency-domain information about spatial displacement. Meanwhile, the motion text embedding E^m supplies semantic guidance for motion-related evolution. These representations are integrated by the motion Prior Fusion module:

$$U^m = \mathcal{F}_m(D^m, P^m, E^m), \tag{3}$$

and then encoded to obtain the motion representation $\widetilde{Z}^m = \mathcal{E}_m(U^m)$.

Intensity Branch. We construct a Statistics Descriptor that summarizes the temporal intensity evolution at each spatial location. Specifically, given the historical radar sequence $X = \{X_l\}_{l=1}^{L}$, we compute the temporal mean and standard deviation as follows:

$$\mu = \frac{1}{L}\sum_{l=1}^{L} X_l, \qquad \sigma = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(X_l - \mu)^2}. \tag{4}$$

To further characterize the tendency of echo intensification or attenuation, we calculate a temporal trend map using a linear temporal slope:

$$\tau = \frac{\sum_{l=1}^{L}(l - \bar{l})(X_l - \mu)}{\sum_{l=1}^{L}(l - \bar{l})^2}, \qquad \bar{l} = \frac{L+1}{2}. \tag{5}$$

These statistics are computed independently at each spatial location, thereby preserving the spatial structure of precipitation systems. We concatenate them and map them into the latent feature space:

$$D^i = \phi_i([\mu \| \sigma \| \tau]), \tag{6}$$

where $\phi_i(\cdot)$ denotes a learnable projection. Here, μ, σ, and τ describe the overall echo strength, temporal fluctuation, and intensity trend, respectively.

The resulting descriptor is further combined with the amplitude prior P^i and the intensity text embedding E^i through the Prior Fusion module:

$$U^i = \mathcal{F}_i(D^i, P^i, E^i), \tag{7}$$

and then encoded to obtain the intensity representation $\widetilde{Z}^i = \mathcal{E}_i(U^i)$.

Dual Cross-Attention Fusion. Since motion and intensity dynamics are strongly coupled, we further introduce dual cross-attention to explicitly model their mutual dependencies. Specifically,

the motion representation attends to the intensity representation as:

$$\widehat{Z}^m = \mathrm{Softmax}\left(\frac{(\widetilde{Z}^m W_Q^m)(\widetilde{Z}^i W_K^i)^\top}{\sqrt{d}}\right)(\widetilde{Z}^i W_V^i), \tag{8}$$

while the reverse interaction is formulated as:

$$\widehat{Z}^i = \mathrm{Softmax}\left(\frac{(\widetilde{Z}^i W_Q^i)(\widetilde{Z}^m W_K^m)^\top}{\sqrt{d}}\right)(\widetilde{Z}^m W_V^m). \tag{9}$$

The enhanced representations are obtained with residual connections: $Z^m = \mathrm{LN}(\widetilde{Z}^m + \widehat{Z}^m)$, $Z^i = \mathrm{LN}(\widetilde{Z}^i + \widehat{Z}^i)$. Finally, the two representations are concatenated and projected into a unified radar dynamics representation:

$$Z^r = \mathcal{P}_r\left([Z^m \| Z^i]\right), \tag{10}$$

which is subsequently fed into the multimodal fusion network.

3.3 Visual Perception Learner

The Visual Perception Learner (VPL) transforms radar observations into visual representations to capture high-level precipitation structures and spatio-temporal patterns beyond radar-native dynamics. VPL first incorporates frequency and periodicity encodings and visual textual priors into the radar representations to enhance temporal and spectral awareness, and then employs spatio-temporal convolution and image normalization to reconstruct radar sequences into image-like representations. Finally, the reconstructed images are fed into a frozen VLM visual encoder to extract visual features, which provide complementary perceptual information for subsequent multimodal fusion.

Frequency and Periodicity Encoding. To capture spectral and temporal dynamics, we apply two complementary encodings to the input radar sequence, explicitly adding frequency-domain and temporal information.

For frequency encoding, we apply the Fast Fourier Transform (FFT) along the temporal dimension:

$$\mathcal{F}_k^t = \sum_{l=1}^{L} X_l \exp\left(-\frac{2\pi j k l}{L}\right), \tag{11}$$

where k denotes the frequency index. The resulting spectral information is further mapped into the frequency representation: $E^f = \mathcal{E}_f(\mathcal{F}^t)$.

For periodicity encoding, we employ sine and cosine functions to explicitly encode temporal dependencies at each time step:

$$E_l^p = \left[\sin\left(\frac{2\pi l}{P}\right), \cos\left(\frac{2\pi l}{P}\right)\right], \tag{12}$$

where P denotes the periodicity parameter. We then aggregate and reshapes these representations: $U = \mathrm{Reshape}(E^r + E^f + E^p)$, where E^r is the original radar encoding.

Spatio-Temporal Convolution. The fused representation is further processed by spatio-temporal convolutions. Specifically, we first apply a 1D convolution along the temporal dimension to capture local temporal dependencies among consecutive radar frames:

$$U^t = \mathrm{Conv1D}_t(U), \tag{13}$$

The resulting features are then processed by 2D convolutions along the spatial dimensions:

$$U^s = \mathrm{Conv2D}_s(U^t), \tag{14}$$

where $\mathrm{Conv2D}_s(\cdot)$ extracts spatial structures within each time step. In this way, the reconstructed representation incorporates both temporal evolution and spatial precipitation patterns.

Image Normalization. The spatio-temporal features are subsequently transformed into image sequences and normalized to match the input range required by the VLM visual encoder. We apply min-max normalization to each reconstructed image:

$$I_{\mathrm{norm}} = 255 \cdot \frac{I_{\mathrm{raw}} - \mathrm{Min}(I_{\mathrm{raw}})}{\mathrm{Max}(I_{\mathrm{raw}}) - \mathrm{Min}(I_{\mathrm{raw}}) + \epsilon}, \tag{15}$$

where ϵ is a small constant for numerical stability. The normalized sequence I_{norm} is then organized into the required image format for visual feature extraction.

VLM visual Encoding. Finally, the reconstructed image sequence is fed into the pre-trained VLM visual encoder:

$$Z^v = \mathcal{E}_{\text{vis}}^{\text{VLM}}(I_{\text{norm}}), \tag{16}$$

where $\mathcal{E}_{\text{vis}}^{\text{VLM}}(\cdot)$ denotes the VLM visual encoder, whose parameters are kept frozen during training. The resulting visual representation Z^v is subsequently fed into the multimodal fusion network.

3.4 Text Prior Generator

To introduce textual priors into radar evolution modeling, we construct a Text Prior Generator (TPG) based on the pre-trained VLM text encoder. Specifically, we design task-related prior prompts covering four aspects of precipitation nowcasting: motion, intensity, visual patterns, and forecasting objectives, with further details provided in Appendix C. The prompts are encoded by the frozen VLM text encoder as follows:

$$E^m, E^i, E^v, E^t = \mathcal{E}_{\text{text}}^{\text{VLM}}\left(Q^m, Q^i, Q^v, Q^t\right), \tag{17}$$

where Q^m, Q^i, Q^v, and Q^t denote the corresponding prior prompts. The resulting motion and intensity embeddings E^m and E^i guide the two radar dynamics branches, while E^v provides semantic guidance for visual perception and E^t supplies task-level priors for subsequent multimodal fusion.

3.5 Multimodal Fusion Network

To integrate complementary representations from different modalities, we construct a Multimodal Fusion Network (MFN) that combines radar-native dynamics, visual representations, and textual semantic priors for precipitation forecasting.

Radar-Vision Fusion. Given the radar dynamics representation Z^r and visual representation Z^v, we concatenate them and employ an MLP to obtain the joint multimodal representation:

$$Z^{rv} = \text{MLP}\left([Z^r \| Z^v]\right), \tag{18}$$

where $[\cdot\|\cdot]$ denotes feature concatenation.

Task-aware Cross-Attention. We further introduce the task-level text embedding E^t to provide an objective-aware prior for forecasting. Specifically, Z^{rv} serves as the query, while E^t is used as the key and value:

$$\widehat{Z}^f = \text{Softmax}\left(\frac{(Z^{rv}W_Q)(E^tW_K)^\top}{\sqrt{d}}\right)(E^tW_V), \tag{19}$$

where W_Q, W_K, and W_V are learnable projection matrices. The final fusion representation is obtained through a residual connection and layer normalization:

$$Z^f = \text{LN}(Z^{rv} + \widehat{Z}^f). \tag{20}$$

Forecasting Predictor. Finally, the fused representation is fed into a forecasting predictor to generate the future radar sequence:

$$\widehat{Y} = \mathcal{P}_{\text{pred}}\left(Z^f\right), \tag{21}$$

where $\widehat{Y} \in \mathbb{R}^{T\times C\times H\times W}$ denotes the predicted radar observations.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on four precipitation benchmark datasets. SEVIR (Veillette et al., 2020) contains 20,393 weather events over the United States, covering 384×384 km at 1 km resolution and 5-minute intervals. MeteoNet (Larvor & Berthomier, 2021) provides over three years of observations from a 550×550 km region in France. Shanghai (Chen et al., 2020) consists of volumetric radar scans collected in Pudong, covering 501×501 km with a 6-minute interval. CIKM (Yao & Li, 2017) includes precipitation events in Guangdong, covering 101×101 km. All radar frames are resized to 128×128 following previous studies. For SEVIR, MeteoNet, and

Table 1: Experiment results on four radar datasets. The best results are highlighted in **bold** and the second-best results are underlined.

	SEVIR						MeteoNet					
	CSI-M↑	CSI-181↑	CSI-219↑	HSS↑	SSIM↑	MSE↓	CSI-M↑	CSI-24↑	CSI-32↑	HSS↑	SSIM↑	MSE↓
ConvGRU	0.2903	0.0879	0.0350	0.3619	0.6100	368.34	0.3401	0.2990	0.1431	0.4667	0.7833	12.85
MAU	0.3076	0.1071	0.0516	0.3863	0.6505	355.48	0.3233	0.2839	0.0997	0.4452	0.7897	12.92
SimVP	0.3108	0.1106	0.0517	0.3924	0.6508	383.56	0.3351	0.3002	0.1130	0.4573	0.7804	13.45
FourCastNet	0.2686	0.0717	0.0339	0.3355	0.5976	410.27	0.3027	0.2533	0.1085	0.4216	0.6450	15.05
Earthformer	0.2892	0.0844	0.0245	0.3665	0.6633	360.11	0.3205	0.2884	0.1237	0.4491	0.7772	14.43
PhyDNet	0.3017	0.1040	0.0278	0.3812	0.6532	357.63	0.3384	0.3194	0.1366	0.4673	0.7823	14.48
EarthFarseer	0.3004	0.0992	0.0413	0.3829	0.6327	388.91	0.3404	0.3170	0.1372	0.4726	0.7542	14.10
NowcastNet	0.2791	0.0770	0.0351	0.3512	0.6839	412.94	0.3427	0.3206	0.1598	0.4751	0.7879	15.64
DiffCast	0.3050	0.1300	0.0582	0.3996	0.6482	559.59	0.3512	0.3340	0.1808	0.4846	0.7887	17.93
AlphaPre	0.3259	0.1332	0.0545	0.4110	0.6884	**345.18**	0.3824	0.3633	0.2002	0.5164	0.7968	12.74
PreMM	**0.3457**	**0.1781**	**0.0950**	**0.4430**	**0.7028**	376.82	**0.4353**	**0.4209**	**0.2574**	**0.5696**	**0.8482**	**8.77**

	Shanghai						CIKM					
	CSI-M↑	CSI-35↑	CSI-40↑	HSS↑	SSIM↑	MSE↓	CSI-M↑	CSI-35↑	CSI-40↑	HSS↑	SSIM↑	MSE↓
ConvGRU	0.3612	0.3163	0.2062	0.4899	0.7796	33.56	0.3091	0.2009	0.1259	0.4006	0.6507	37.13
MAU	0.3983	0.3621	0.2417	0.5346	0.7195	30.40	0.3039	0.2054	0.1241	0.3928	0.6325	40.74
SimVP	0.3850	0.3549	0.2382	0.5194	0.7795	34.40	0.3052	0.2044	0.1321	0.3955	0.6538	38.06
FourCastNet	0.3571	0.3108	0.2073	0.4868	0.5598	32.10	0.2980	0.1849	0.1015	0.3801	0.4359	36.14
Earthformer	0.3503	0.3178	0.1872	0.4844	0.7298	35.57	0.3077	0.2039	0.1369	0.4001	0.6267	36.49
PhyDNet	0.3654	0.3236	0.2176	0.4957	0.7751	36.41	0.3038	0.2052	0.1287	0.3931	0.6541	39.56
EarthFarseer	0.3926	0.3608	0.2343	0.5330	0.5405	32.68	0.3000	0.2046	0.1259	0.3911	0.6373	39.87
NowcastNet	0.3953	0.3608	0.2450	0.5334	0.7902	33.56	0.2991	0.1940	0.1188	0.3865	0.6713	40.96
DiffCast	0.4089	0.3740	0.2606	0.5476	0.7879	36.35	0.3159	0.2009	0.1457	0.4085	0.6499	42.78
AlphaPre	0.4178	0.3854	0.2615	0.5534	0.7951	28.02	0.3194	0.2068	0.1416	0.4137	0.6568	35.18
PreMM	**0.4429**	**0.4108**	**0.3086**	**0.5803**	**0.8072**	**22.98**	**0.3231**	**0.2243**	**0.1528**	**0.4160**	**0.6785**	**35.02**

Shanghai, we use 5 input frames to predict the next 20 frames, while for CIKM the input and output frames are set to 5 and 10, respectively. More detailed dataset statistics and preprocessing settings are provided in the appendix.

Evaluation. We evaluate our model using four widely used metrics: Critical Success Index (CSI), Heidke Skill Score (HSS), Structural Similarity Index Measure (SSIM), and Mean Squared Error (MSE). CSI measures the accuracy of precipitation event detection by considering both false alarms and missed events. Following previous works (Yu et al., 2024; Lin et al., 2025), we calculate CSI at multiple thresholds: $\{16, 74, 133, 160, 181, 219\}$ for SEVIR, $\{12, 18, 24, 32\}$ for MeteoNet, and $\{20, 30, 35, 40\}$ for Shanghai and CIKM, and report their average value (CSI-M). HSS evaluates the forecasting skill against random predictions, while SSIM and MSE measure the structural similarity and pixel-level error between predicted and observed radar sequences.

Baselines. We compare our method against ten representative precipitation nowcasting baselines, including ConvGRU (Shi et al., 2017a), MAU (Chang et al., 2021), SimVP (Gao et al., 2022a), FourCastNet (Pathak et al., 2022), EarthFormer (Gao et al., 2022c), PhyDNet (Guen & Thome, 2020), EarthFarseer (Wu et al., 2024), NowcastNet (Zhang et al., 2023), DiffCast (Yu et al., 2024), and Alphapre (Lin et al., 2025). For fairness, all baseline results are obtained using the official implementations released by the original authors.

4.2 Main Results

Table 1 reports the comparison of PreMM with representative baselines on four radar datasets. Overall, PreMM achieves the best performance on 23 out of 24 evaluation settings, demonstrating consistent advantages across different precipitation scenarios. More importantly, PreMM exhibits clear advantages at higher CSI thresholds, which better reflect the forecasting performance for heavy precipitation. On SEVIR, PreMM improves CSI-181 and CSI-219 over the best competing results by 33.7% and 63.2%, respectively. Consistent improvements are also observed on MeteoNet and Shanghai, where the gains become more pronounced at higher thresholds. These results suggest that incorporating visual representations and semantic priors enables PreMM to better capture complex precipitation evolution, particularly for intense rainfall patterns. Figure 2 further compares the CSI and HSS scores of different methods across forecasting time steps. Our method maintains consistently superior performance throughout the forecasting horizon, achieving the best results at nearly all time steps. Figure 3 presents a representative forecasting example. Compared with other meth-

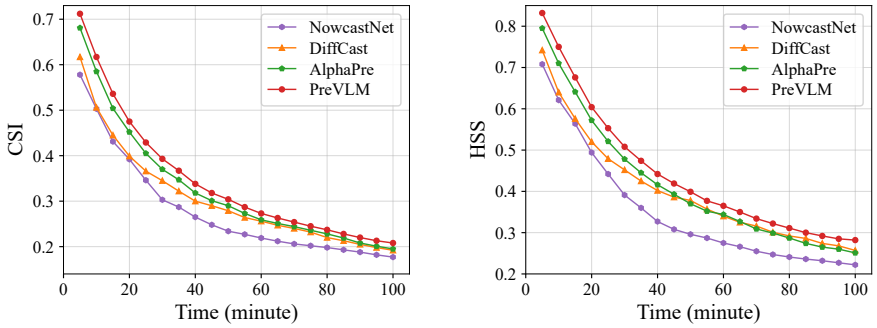

Figure 2: The CSI and HSS performance at different forecasting time steps on the SEVIR dataset.

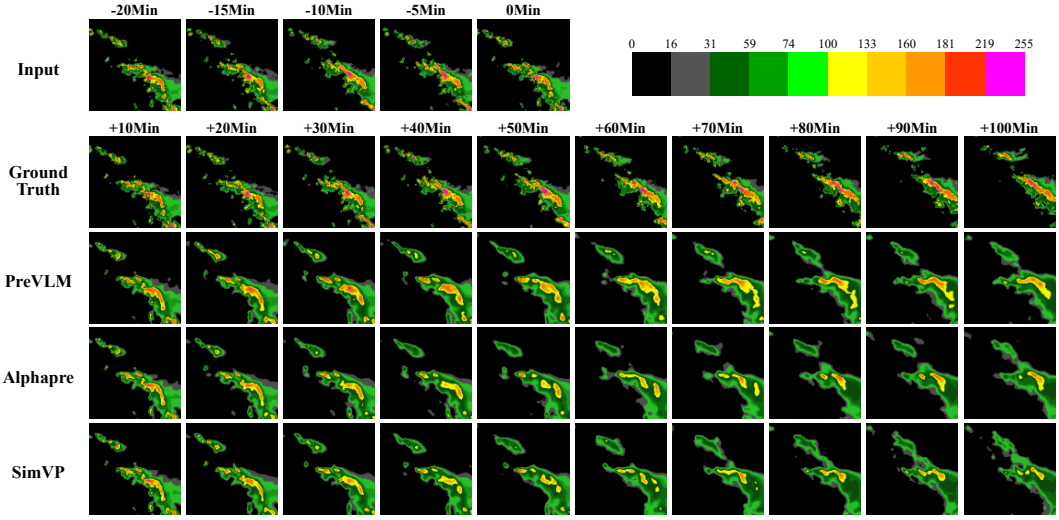

Figure 3: A visualization example from the SEVIR dataset.

ods, which tend to underestimate precipitation intensity at later forecasting steps, our model better preserves the intensity evolution and produces predictions closer to the ground truth.

4.3 Ablation Study

PreMM consists of three key components: the Radar Dynamics Learner (RDL), the Visual Perception Learner (VPL), and the Text Prior Generator (TPG). To evaluate the effectiveness of these components, we conduct several ablation experiments. Table 2 presents the results of these variants. Removing any component leads to consistent performance degradation, demonstrating the complementary contributions of different modalities. In particular, removing RDL causes the largest performance drop, highlighting the importance of explicitly modeling motion and intensity dynamics in radar evolution. The degradation caused by removing VPL further verifies that visual representations provide complementary structural information beyond radar signals. Meanwhile, the performance decrease of removing TPG confirms that semantic priors from the VLM can provide effective guidance for precipitation forecasting.

Table 2: The ablation study results for the components of PreMM.

	Shanghai						MeteoNet					
	CSI-M↑	CSI-35↑	CSI-40↑	HSS↑	SSIM↑	MSE↓	CSI-M↑	CSI-24↑	CSI-32↑	HSS↑	SSIM↑	MSE↓
w/o RDL	0.4019	0.3508	0.2196	0.5287	0.7785	29.15	0.3928	0.3597	0.1915	0.5123	0.8087	11.38
w/o VPL	0.4168	0.3725	0.2479	0.5486	0.7864	27.43	0.4057	0.3776	0.2142	0.5298	0.8196	10.62
w/o TPG	0.4326	0.3975	0.2917	0.5674	0.7998	24.36	0.4212	0.4027	0.2405	0.5518	0.8375	9.34
PreMM	**0.4429**	**0.4108**	**0.3086**	**0.5803**	**0.8072**	**22.98**	**0.4353**	**0.4209**	**0.2574**	**0.5696**	**0.8482**	**8.77**

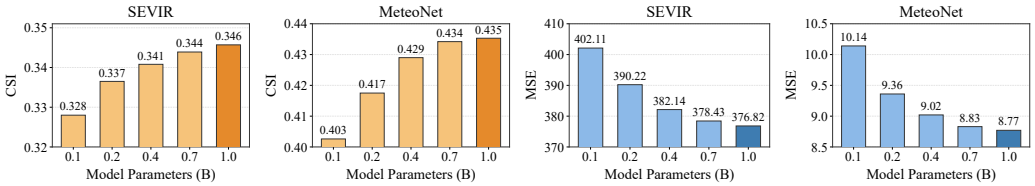

Figure 4: CSI and MSE with varying model parameters on the SEVIR and MeteoNet datasets.

Table 3: Comparison of different VLM backbones on the SEVIR dataset.

VLM Type	Params	CSI-M↑	CSI-181↑	CSI-219↑	HSS↑	SSIM↑	MSE↓
CLIP	150M	0.3413	0.1742	0.0896	0.4385	0.7013	384.68
BLIP	500M	0.3287	0.1672	0.0859	0.4293	0.6957	382.36
InternVL2.5	1B	0.3341	0.1706	0.0878	0.4339	0.6974	381.65
Qwen2.5-VL	3B	**0.3457**	**0.1781**	**0.0950**	**0.4430**	**0.7028**	**376.82**

4.4 Effect of Model Size on Performance

We investigate the scalability of PreMM by varying its model size from 0.1B to 1.0B parameters. As shown in Figure 4, increasing model capacity consistently improves performance on both SEVIR and MeteoNet, yielding higher CSI and lower MSE. Notably, the gains are more pronounced when scaling from smaller models, indicating that additional capacity substantially strengthens the modeling of complex precipitation structures and dynamics. The highly consistent trends across two datasets further indicate that the benefit of scaling is not dataset-specific, but reflects the ability of PreMM to better exploit multimodal representations with increased capacity. Overall, these results demonstrate the favorable and robust scalability of our framework.

4.5 Effect of VLM Type on Performance

We further investigate the impact of different pre-trained VLMs by replacing the VLM backbone while keeping the remaining architecture unchanged. As shown in Table 3, PreMM achieves competitive performance across all VLM variants, demonstrating its compatibility with different vision-language representations. Among them, Qwen2.5-VL consistently achieves the best performance across all metrics. Compared with the best competing VLM variant, it improves CSI-M, CSI-181, and CSI-219 by 1.3%, 2.2%, and 6.0%, respectively. Notably, performance does not strictly increase with the parameter size of the VLM, as the lightweight CLIP variant outperforms the larger BLIP and InternVL2.5 on several CSI-based metrics. This indicates that the effectiveness of the VLM depends not only on model scale, but also on the suitability of its learned visual-semantic representations for radar precipitation patterns. Overall, these results demonstrate that PreMM is robust to different VLM choices, while stronger vision-language representations can further improve precipitation forecasting performance.

5 Conclusion

In this paper, we propose PreMM, a multimodal precipitation nowcasting framework that unifies radar-native dynamics, visual representations, and textual semantic priors. PreMM introduces a Radar Dynamics Learner (RDL) to explicitly capture motion and intensity dynamics, and enhances them with frequency-domain and textual priors; a Visual Perception Learner (VPL) to reconstruct radar observations into visual representations for capturing high-level precipitation structures; a Text Prior Generator (TPG) to provide meteorological semantic knowledge for multimodal learning. A Multimodal Fusion Network (MFN) is further designed to effectively integrate these complementary representations for future radar prediction. Extensive experiments on four benchmark datasets demonstrate that PreMM consistently outperforms existing state-of-the-art methods, especially for heavy precipitation prediction with high-intensity thresholds. The ablation studies further validate the effectiveness of RDL, VPL, and TPG. These results highlight the potential of multimodal representation learning for precipitation nowcasting.

REFERENCES

Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Yan Ye, Xiang Xinguang, and Wen Gao. Mau: A motion-aware unit for video prediction and beyond. In *Advances in Neural Information Processing Systems*, 2021.

Jian Chen, Peilin Zhou, Yining Hua, Dading Chong, Meng Cao, Yaowei Li, Wei Chen, Bing Zhu, Junwei Liang, and Zixuan Yuan. Climateiqa: A new dataset and benchmark to advance vision-language models in meteorology anomalies analysis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025.

Lei Chen, Yuan Cao, Leiming Ma, and Junping Zhang. A deep learning-based methodology for precipitation nowcasting with radar. *Earth and Space Science*, 7:e2019EA000812, 2020.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022b.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. In *Advances in Neural Information Processing Systems*, 2022c.

Junchao Gong, Siwei Tu, Weidong Yang, Ben Fei, Kun Chen, Wenlong Zhang, Xiaokang Yang, Wanli Ouyang, and Lei Bai. Postcast: Generalizable postprocessing for precipitation nowcasting via unsupervised blurriness modeling. In *International Conference on Learning Representations*, 2025.

Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Gwennaëlle Larvor and Lea Berthomier. Meteonet: An open reference weather dataset for ai by météo-france. In *American Meteorological Society Meeting Abstracts*, 2021.

Dawei Li, Kefeng Deng, Di Zhang, Yudi Liu, Hongze Leng, Fukang Yin, Kaijun Ren, and Junqiang Song. Lpt-qpn: A lightweight physics-informed transformer for quantitative precipitation nowcasting. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–19, 2023.

Haobo Li, Zhaowei Wang, Jiachen Wang, YueYa Wang, Alexis Kai Hon Lau, and Huamin Qu. Cllmate: A multimodal benchmark for weather and climate events forecasting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Wenhui Li, Ying Zhou, Yue Li, Dan Song, Zhiqiang Wei, and An-An Liu. Hierarchical transformer with lightweight attention for radar-based precipitation nowcasting. *IEEE Geoscience and Remote Sensing Letters*, 21:1–5, 2024.

Kenghong Lin, Baoquan Zhang, Demin Yu, Wenzhi Feng, Shidong Chen, Feifan Gao, Xutao Li, and Yunming Ye. Alphapre: Amplitude-phase disentanglement model for precipitation nowcasting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Xudong Ling, Chaorong Li, Tianxi Huang, Qian Dong, and Guiduo Duan. Langprecip: Language-aware multimodal precipitation nowcasting. In *Forty-third International Conference on Machine Learning*, 2026.

Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. Weatherqa: Can multimodal language models reason about severe weather? In *arXiv preprint arXiv:2406.11217*, 2024.

Xiang Pan, Yinghui Lu, Hao Huang, Gang Chen, Ang Zhou, and Kun Zhao. Advancing convective precipitation nowcasting via 3d polarimetric radar data and physics-constrained deep learning model. *Geophysical Research Letters*, 53:e2025GL120431, 2026.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. In *arXiv preprint arXiv:2202.11214*, 2022.

Bernardo Perrone Ribeiro and Jana Faganeli Pucer. Flowcast: Advancing precipitation nowcasting with conditional flow matching. In *International Conference on Learning Representations*, 2026.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597:672–677, 2021.

Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. *Advances in neural information processing systems*, 36:45259–45287, 2023.

Mahdi Gholami Sharafkhane, Ali Naghi Ziaei, Seyed Mohammadreza Naghedifar, Amir Akbari, and Amir Verdi. Aquacrop plug-in-pso: A novel irrigation scheduling optimization framework for maize to maximize crop water productivity using in-season weather forecast and crop yield estimation. *Agricultural Water Management*, 306:109153, 2024.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, 2015.

Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. In *Advances in neural information processing systems*, 2017a.

Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. In *Advances in neural information processing systems*, 2017b.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In *Advances in neural information processing systems*, 2020.

Shijin Wang, Min Sun, and Yinglin Li. Nowcasting echo top for aviation operations using cnn-transformer. *IEEE Transactions on Intelligent Transportation Systems*, 26:21724–21733, 2025.

Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S Yu. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

Binqing Wu, Weiqi Chen, Shiyu Liu, Zongjiang Shang, Haiou Wang, Liang Sun, and Ling Chen. Mocast: Learning turbulent motions under physical guidance for precipitation nowcasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Hao Wu, Yuxuan Liang, Wei Xiong, Zhengyang Zhou, Wei Huang, Shilong Wang, and Kun Wang. Earthfarsser: Versatile spatio-temporal dynamical systems modeling in one model. In *Proceedings of the AAAI conference on artificial intelligence*, 2024.

Yichen Yao and Zhongjie Li. Cikm analyticup 2017: short-term precipitation forecasting based on radar reflectivity images. In *Proceedings of the Conference on Information and Knowledge Management, Short-Term Quantitative Precipitation Forecasting Challenge, Singapore*, 2017.

Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yang Zeng, Minghua Hu, Haiyan Chen, Ligang Yuan, Sameer Alam, and Dabin Xue. Improved air traffic flow prediction in terminal areas using a multimodal spatial–temporal network for weather-aware (mst-wa) model. *Advanced Engineering Informatics*, 62:102935, 2024.

Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619: 526–532, 2023.

Zheng Zhang, Chuyao Luo, Shanshan Feng, Rui Ye, Yunming Ye, and Xutao Li. Rap-net: Region attention predictive network for precipitation nowcasting. *Geoscientific Model Development*, 15: 5407–5419, 2022.

A Experimental Details

A.1 Dataset Descriptions

We conduct experiments on four widely used radar-based precipitation nowcasting benchmarks, including SEVIR, Shanghai Radar, MeteoNet, and CIKM. These datasets cover geographically diverse regions and different meteorological conditions, providing a comprehensive testbed for evaluating the spatiotemporal modeling capability of nowcasting models.

SEVIR. The Storm EVent ImagRy (SEVIR) dataset (Veillette et al., 2020) is a large-scale multi-sensor weather dataset collected over the contiguous United States (CONUS). In this work, we use the Vertically Integrated Liquid (VIL) product derived from NEXRAD radar observations. SEVIR contains more than 10,000 weather events, where each event covers a spatial domain of approximately $384, \mathrm{km} \times 384, \mathrm{km}$ and is represented by radar maps of size 384×384. Each event spans four hours with observations sampled at 5-minute intervals. Following the commonly adopted temporal split, events from 2017–2018 are used for training, while those from 2019 are reserved for validation and testing.

Shanghai. The Shanghai Radar dataset (Chen et al., 2020) consists of composite reflectivity (CR) observations acquired by a dual-polarization Doppler weather radar located in Pudong, Shanghai. The dataset spans from October 2015 to July 2018, with radar volume scans produced at approximately 6-minute intervals. The original observations are interpolated onto a regular longitude–latitude grid with a spacing of $0.01^\circ \times 0.01^\circ$, resulting in radar maps of size 501×501. We use observations from 2015 to 2017 for model training and reserve the data collected in 2018 for validation and testing.

MeteoNet. MeteoNet (Larvor & Berthomier, 2021) is an open meteorological dataset released by Météo-France and covers two geographical regions in north-western and south-eastern France. The dataset contains radar observations collected between 2016 and 2018, with radar scans available every 5 minutes on a 0.01° spatial grid. Following the standard nowcasting benchmark configuration, we extract a 400×400 sub-region from the original radar fields. Data from 2016 and 2017 are used for training, whereas observations from 2018 are used for validation and testing.

CIKM. The CIKM radar dataset (Yao & Li, 2017) was released for the CIKM AnalytiCup 2017 precipitation forecasting competition. It contains radar echo sequences collected over Guangdong, China, with each frame covering an approximately $101, \mathrm{km} \times 101, \mathrm{km}$ region. Radar observations are recorded at 6-minute intervals, and each original sequence consists of 15 consecutive radar maps. Accordingly, we use the first five frames as observations and predict the subsequent ten frames.

To eliminate differences in native spatial dimensions and enable a unified model configuration across datasets, all radar frames are resized to 128×128 before being fed into the model. For SEVIR, Shanghai Radar, and MeteoNet, the input and output sequence lengths are set to 5 and 20. For CIKM, the input and output sequence lengths are set to 5 and 10. Unless otherwise specified, the same preprocessing and sequence construction protocols are applied to all compared methods to ensure a fair evaluation.

A.2 IMPLEMENTATION

We mainly follow the experimental configurations in (Lin et al., 2025) across all baselines within a unified evaluation pipeline for fair comparisons.

Our model is trained for 100 epochs using the Adam optimizer, with a batch size of 4 and an initial learning rate of 1×10^{-4}. We use CLIP as the default vision-language model (VLM) and keep its parameters frozen during training. The input radar observations are represented as three-channel images and the periodicity parameter is set to 5. The main model uses a hidden dimension of 512 with a dropout rate of 0.1, while the multimodal fusion module projects features into a 64-dimensional space. For the radar branch, we employ a dual cross-attention mechanism with 4 attention heads and a hidden dimension of 64. The radar inputs are processed using a patch size of 4 and a backbone stride of 2, with a local attention window of 8. We adopt a drop-path rate of 0.15 and enable gradient checkpointing to reduce memory consumption during training. The auxiliary prediction objective is introduced with a loss weight of 0.2. We further set the normalization constant to 0.4 to balance the feature representations across different modalities. All experiments are repeated three times and we report the average results.

A.3 EVALUATION METRICS

We assess the precipitation nowcasting models using both event-based and continuous-field metrics.

Event-based metrics. For a given rainfall intensity threshold τ (e.g., $\tau = 30$ mm/h), we binarize the ground truth and prediction into "heavy precipitation" events. Let TP, FP, FN, and TN denote the numbers of true positives, false positives, false negatives, and true negatives, respectively.

Critical Success Index (CSI). CSI is defined as

$$\mathrm{CSI} = \frac{TP}{TP + FP + FN}, \tag{22}$$

with $0 \leq \mathrm{CSI} \leq 1$. Higher CSI indicates better performance. CSI ignores true negatives (TN), which is desirable because in most radar images the majority of pixels are non-event; including TN would inflate accuracy even for trivial "no rain" predictions.

CSI-M. CSI-M is the average of CSI values computed over multiple thresholds τ. Lower thresholds capture moderate precipitation, while higher thresholds focus on extreme events. This averaged score provides a more comprehensive assessment across precipitation intensities.

Heidke Skill Score (HSS). HSS measures the improvement of the forecast over a random (chance) forecast. It is computed as

$$\mathrm{HSS} = \frac{2(TP \cdot TN - FP \cdot FN)}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)}. \tag{23}$$

HSS ranges from $-\infty$ to 1, with 1 indicating perfect skill and 0 corresponding to no skill relative to random guessing.

Continuous-field metrics. Unlike threshold-based scores, the following metrics operate directly on the original radar reflectivity or rainfall intensity maps without binarization.

Mean Squared Error (MSE). MSE is the pixel-wise average squared difference between the prediction y and the ground truth x:

$$\mathrm{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i)^2. \tag{24}$$

It penalises large deviations but does not explicitly capture spatial structure.

Structural Similarity Index (SSIM). SSIM assesses the perceptual similarity between two images by comparing luminance, contrast, and structure. For two images x and y, SSIM is defined as

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{25}$$

Table 4: An overview of the experimental configurations for PreMM.

Dataset	Model Hyperparameter				Training Process		
	Input Length	Output Length	Patch	Heads	LR*	Batch Size	Epochs
SEVIR	5	20	4	8	10^{-3}	2	100
MeteoNet	5	20	8	8	10^{-3}	2	100
Shanghai	5	20	4	8	10^{-3}	2	100
CIKM	5	10	4	8	10^{-3}	4	100

where μ_x, μ_y are the mean intensities, σ_x, σ_y are the standard deviations, σ_{xy} is the covariance, and C_1, C_2 are small constants for numerical stability. SSIM ranges from -1 to 1 (typically 0 to 1), with higher values indicating better structural fidelity. It is particularly sensitive to the shape, texture, and spatial distribution of precipitation patterns.

Throughout our experiments, we report CSI, CSI-M, HSS, MSE, and SSIM to provide a holistic evaluation of both event detection accuracy and continuous field reconstruction quality.

A.4 Model Configuration

The model configurations for different datasets are summarized in Table 4. By default, the Adam optimizer (Kingma & Ba., 2015) is employed throughout all experiments. The input length is fixed to 5 historical radar frames for all datasets, while the forecasting horizon is set to 20 frames for SEVIR, MeteoNet, and Shanghai, and 10 frames for CIKM. The patch size is set to 4 by default, except for MeteoNet, where a patch size of 8 is adopted. We employ 8 attention heads across all datasets. For model optimization, the learning rate is fixed at 1×10^{-3}, and all models are trained for 100 epochs. The batch size is set to 2 for SEVIR, MeteoNet, and Shanghai, and increased to 4 for CIKM. Moreover, we utilize the Qwen2.5-VL 3B model for all datasets. These configurations are kept consistent unless otherwise specified in the corresponding experiments.

B MODEL WORKFLOW

Workflow 1 Multimodal Radar Precipitation Forecasting

Require: Historical radar sequence $X = \{X_l\}_{l=1}^{L}$, Radar Dynamics Learner $\mathcal{R}$, Visual Perception Learner $\mathcal{V}$, Text Encoder $\mathcal{E}_{\text{text}}^{\text{VLM}}$, Visual Encoder $\mathcal{E}_{\text{vis}}^{\text{VLM}}$.
Ensure: Future radar prediction $\hat{Y}$.

1: Generate semantic priors:

$$(E^m, E^i, E^v, E^t) \leftarrow \mathcal{E}_{\text{text}}^{\text{VLM}}(Q^m, Q^i, Q^v, Q^t)$$

2: Decompose radar dynamics into motion and intensity cues:

$$D_l^m = X_l - X_{l-1}, \qquad D^i = \phi_i([\mu \| \sigma \| \tau])$$

3: Extract spectral priors from radar observations:

$$P^m = \mathcal{G}_p(\text{Phase}(\text{FFT}(X))), \quad P^i = \mathcal{G}_a(\text{Amp}(\text{FFT}(X)))$$

4: Encode radar dynamics with motion-intensity interaction:

$$Z^r \leftarrow \mathcal{R}(D^m, D^i, P^m, P^i, E^m, E^i)$$

5: Generate visual-enhanced radar representation:

$$U \leftarrow \mathcal{V}(X, \mathcal{F}^t, E^p)$$

6: Reconstruct normalized image sequence:

$$I_{\text{norm}} \leftarrow 255 \frac{U - \min(U)}{\max(U) - \min(U) + \epsilon}$$

7: Extract visual feature:

$$Z^v \leftarrow \mathcal{E}_{\text{vis}}^{\text{VLM}}(I_{\text{norm}})$$

8: Fuse radar and visual representations:

$$Z^{rv} \leftarrow \text{MLP}([Z^r \| Z^v])$$

9: Perform task-aware semantic fusion:

$$Z^f \leftarrow \text{LN}(Z^{rv} + \text{CA}(Z^{rv}, E^t))$$

10: Predict future precipitation:

$$\hat{Y} \leftarrow \mathcal{P}_{\text{pred}}(Z^f)$$

11: **return** $\hat{Y}$

C TEXTUAL PRIOR CONSTRUCTION

To provide explicit semantic knowledge for different components of PreMM, we construct four meteorology-oriented prompts corresponding to motion dynamics, intensity dynamics, visual patterns, and the precipitation nowcasting objective. Instead of using a generic description of radar observations, each prompt is designed to emphasize a distinct aspect of radar echo evolution that is consistent with the corresponding modeling branch. Specifically, the motion prompt focuses on displacement and structural transport, the intensity prompt describes the strengthening and weakening of radar echoes, the visual prompt characterizes high-level spatial morphology and organization, and the task prompt encodes forecasting-oriented knowledge about future precipitation evolution. The complete prompts are provided below.

Motion Prior Prompt Q^m

Prompt: Act as a professional radar meteorology analyst and characterize the motion dynamics of precipitation echoes from a radar sequence. Focus on how echo structures move and geometrically evolve over time.
Analysis:

1. **Displacement:** Identify the dominant propagation direction, relative speed, and spatial displacement.
2. **Motion Pattern:** Describe translation, rotation, expansion, contraction, or deformation.
3. **Structural Evolution:** Identify merging, splitting, stretching, and displacement of major echo regions.
4. **Temporal Consistency:** Determine whether motion tendencies are persistent or change over time.

Requirements: Provide a concise description emphasizing motion and structural transport; avoid unnecessary discussion of echo intensity or unsupported meteorological phenomena.

Intensity Prior Prompt Q^i

Prompt: Act as a professional radar meteorology analyst and characterize the intensity evolution of precipitation echoes. Focus on temporal changes in echo strength rather than spatial displacement.
Analysis:

1. **Mean State:** Describe the persistent strength and spatial distribution of weak and strong echoes.
2. **Variability:** Identify stable regions and areas with large temporal intensity fluctuations.
3. **Evolution Trend:** Determine whether echoes strengthen, weaken, emerge, maintain, or dissipate.
4. **Spatial Variation:** Locate regions where intensification or attenuation mainly occurs.

Requirements: Summarize the dominant intensity state, variability, and strengthening or weakening tendency using concise meteorological terminology.

Visual Prior Prompt Q^v

Prompt: Act as a professional radar meteorology analyst and characterize the high-level visual patterns of a radar echo sequence, focusing on precipitation morphology and spatial organization.
Analysis:

1. **Spatial Distribution:** Describe the locations, concentration, connectivity, and arrangement of major echo regions.
2. **Morphology:** Identify cellular, banded, linear, arc-shaped, irregular, or other dominant structures.
3. **Boundary and Structure:** Describe echo boundaries, internal cores, weak-echo gaps, and multi-scale organization.
4. **Visual Evolution:** Characterize changes in morphology, organization, and structural complexity over time.

Requirements: Provide a concise visual description emphasizing high-level structural patterns that complement radar-native motion and intensity dynamics.

Task Prior Prompt Q^t

Prompt: Act as an expert in radar-based precipitation nowcasting and describe the forecasting knowledge required to predict future radar echoes from historical observations.
Forecasting:

1. **Motion Extrapolation:** Extend persistent displacement, deformation, merging, and splitting tendencies into the future.
2. **Intensity Extrapolation:** Consider persistent strengthening, weakening, emergence, and dissipation trends.
3. **Structural Continuity:** Preserve coherent morphology, connectivity, and spatial organization when supported by observations.
4. **Coupled Evolution:** Jointly consider motion, intensity, visual structure, and temporal consistency for future prediction.

Requirements: Summarize the key principles for generating a temporally coherent and observation-consistent future radar sequence without introducing unsupported structures or abrupt changes.

D HYPERPARAMETER SENSITIVITY

We investigate the sensitivity of PreMM to the text prior fusion weight λ_t, which controls the contribution of semantic guidance in the task-aware cross-attention module. Specifically, we modify the fusion operation as:

$$Z^f = \mathrm{LN}(Z^{rv} + \lambda_t \widehat{Z}^f),$$

and evaluate different values of λ_t on four benchmark datasets. The results are shown in Table 5.

Table 5: Sensitivity analysis of the text prior fusion weight λ_t. The default setting is $\lambda_t = 1.0$.

λ_t	SEVIR				MeteoNet			
	CSI-M↑	CSI-219↑	HSS↑	MSE↓	CSI-M↑	CSI-32↑	HSS↑	MSE↓
0.25	0.3401	0.0918	0.4381	382.65	0.4297	0.2528	0.5618	9.02
0.5	0.3442	0.0941	0.4421	378.95	0.4338	0.2568	0.5675	8.82
1.0	**0.3457**	**0.0950**	**0.4430**	**376.82**	**0.4353**	0.2574	**0.5696**	**8.77**
2.0	0.3438	0.0936	0.4412	379.84	0.4329	**0.2581**	0.5669	8.91
λ_t	Shanghai				CIKM			
	CSI-M↑	CSI-40↑	HSS↑	MSE↓	CSI-M↑	CSI-40↑	HSS↑	MSE↓
0.25	0.4372	0.3015	0.5724	23.21	0.3187	0.1486	0.4118	36.94
0.5	0.4406	0.3068	0.5778	23.05	**0.3234**	**0.1534**	0.4154	**35.02**
1.0	**0.4429**	**0.3086**	**0.5803**	**22.98**	0.3231	0.1528	**0.4160**	35.47
2.0	0.4394	0.3041	0.5748	23.42	0.3204	0.1502	0.4129	36.79

We observe that PreMM maintains stable performance across different values of λ_t, demonstrating the robustness of semantic prior integration. As λ_t increases, performance generally improves and reaches a favorable balance around $\lambda_t = 1.0$. However, the optimal value slightly varies across datasets. For example, MeteoNet achieves the best CSI-32 with $\lambda_t = 2.0$, while CIKM obtains its best CSI-related metrics with $\lambda_t = 0.5$. This indicates that different precipitation scenarios may require different levels of semantic guidance. Excessively large fusion weights may introduce semantic bias and weaken radar-specific dynamics, leading to performance degradation. Therefore, we adopt $\lambda_t = 1.0$ as the default setting, which provides consistently strong performance across diverse radar datasets.

E Additional Results

E.1 Additional Evaluation Matrices

To provide a more comprehensive evaluation of PreMM, we further report additional experimental results on four precipitation nowcasting benchmarks with more evaluation metrics, including MAE and LPIPS, as shown in Table 6. Compared with the main results, these metrics provide complementary perspectives for assessing forecasting quality. MAE measures the average absolute pixel-level deviation between predicted and ground-truth radar frames, which can reflect the overall intensity reconstruction accuracy. In addition, we introduce the Learned Perceptual Image Patch Similarity (LPIPS) to evaluate the perceptual similarity between predicted and observed precipitation patterns. Different from conventional pixel-wise metrics, LPIPS compares deep feature representations extracted from a pretrained network and measures the perceptual distance between two images. Specifically, given the predicted frame $\hat{Y}$ and the ground-truth frame Y, LPIPS is defined as:

$$\mathrm{LPIPS}(\hat{Y}, Y) = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \left\| \omega_l \odot (\phi_l(\hat{Y})_{hw} - \phi_l(Y)_{hw}) \right\|_2^2, \tag{26}$$

where $\phi_l(\cdot)$ denotes the feature representation extracted from the l-th layer of a pretrained network, ω_l is a learnable channel-wise scaling factor, and H_l and W_l represent the spatial dimensions of the corresponding feature map. A lower LPIPS value indicates higher perceptual consistency between the predicted and ground-truth precipitation structures.

As shown in Table 6, PreMM achieves consistently strong performance across all datasets and evaluation metrics. In particular, the low LPIPS scores demonstrate that our predictions preserve not only pixel-level accuracy but also high-level precipitation structures and spatial patterns. Together with CSI, HSS, SSIM, MSE, and MAE results, these additional evaluations further verify the effectiveness of PreMM in capturing complex precipitation evolution from multiple perspectives.

Table 6: Overall experimental results of PreMM on four precipitation nowcasting benchmarks.

Dataset	Evaluation Metrics									
SEVIR	CSI-M↑	CSI-133↑	CSI-160↑	CSI-181↑	CSI-219↑	HSS↑	SSIM↑	MSE↓	MAE↓	LPIPS↓
	0.3457	0.3124	0.2094	0.1781	0.0950	0.4430	0.7028	376.82	8.63	0.2797
MeteoNet	CSI-M↑	CSI-12↑	CSI-18↑	CSI-24↑	CSI-32↑	HSS↑	SSIM↑	MSE↓	MAE↓	LPIPS↓
	0.4353	0.5584	0.5046	0.4209	0.2574	0.5696	0.8482	8.77	1.04	0.1755
Shanghai	CSI-M↑	CSI-20↑	CSI-30↑	CSI-35↑	CSI-40↑	HSS↑	SSIM↑	MSE↓	MAE↓	LPIPS↓
	0.4429	0.5632	0.4856	0.4108	0.3086	0.5803	0.8072	22.98	1.63	0.2483
CIKM	CSI-M↑	CSI-20↑	CSI-30↑	CSI-35↑	CSI-40↑	HSS↑	SSIM↑	MSE↓	MAE↓	LPIPS↓
	0.3231	0.6237	0.2916	0.2243	0.1528	0.4160	0.6785	35.02	3.01	0.3815

E.2 Additional Comparison Results

To further examine the performance of PreMM under different precipitation intensities, we provide a detailed comparison with AlphaPre across all CSI thresholds on the four datasets in Figure 5. While CSI-M summarizes the average performance over multiple thresholds, the threshold-wise results provide a finer-grained view of the forecasting capability for precipitation events with different intensities. Overall, PreMM outperforms AlphaPre at nearly all thresholds across the four datasets. More importantly, the advantage becomes increasingly pronounced at higher thresholds, which correspond to more intense precipitation events. For example, on SEVIR, the performance gap increases from only marginal differences at low thresholds to gains of 0.045 and 0.041 at thresholds 181 and 219, respectively. Similar trends can be observed on MeteoNet and Shanghai, where PreMM improves the CSI by 0.057 at threshold 32 and by 0.047 at threshold 40, respectively. On

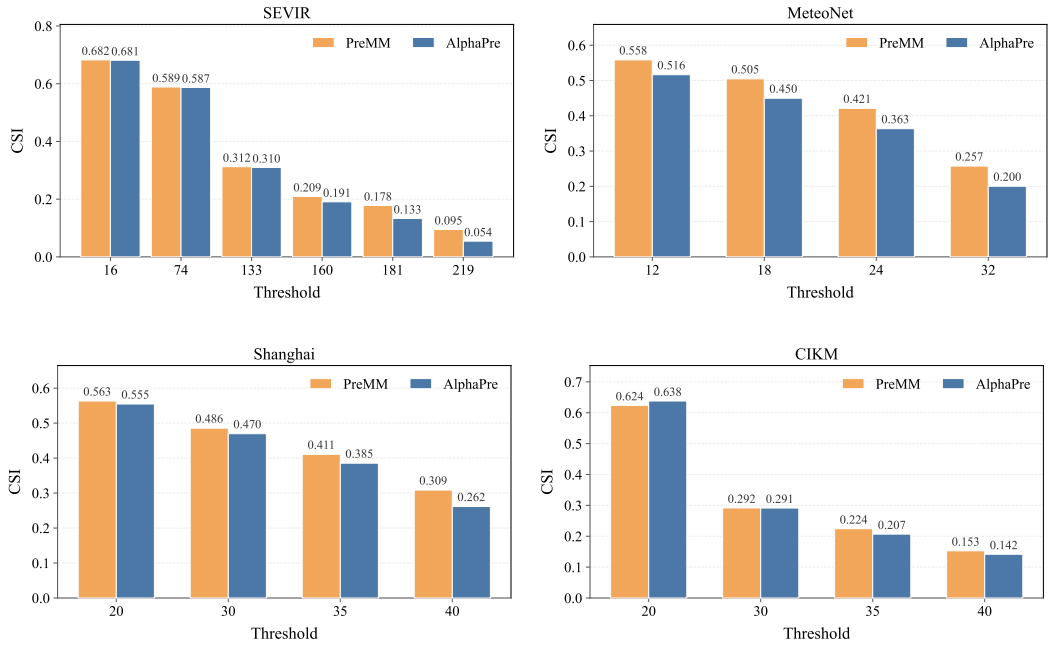

Figure 5: The comparison results of all threshold on four datasets

CIKM, although AlphaPre performs slightly better at the lowest threshold, PreMM consistently surpasses it at the remaining thresholds and maintains a clear advantage for stronger precipitation. These results further demonstrate that the proposed multi-perspective modeling framework is particularly effective in capturing the complex evolution of high-intensity precipitation, while maintaining competitive performance across a broad range of rainfall intensities.

F Future Work

Although PreMM demonstrates the effectiveness of jointly modeling radar-native dynamics, visual representations, and textual semantic priors, several promising directions remain for further exploration.

Dynamic and Event-specific Textual Priors. The textual priors in the current framework are constructed from predefined meteorological prompts and encoded by a frozen VLM. While such priors provide stable semantic guidance, they are shared across different radar sequences and thus cannot fully reflect event-specific precipitation evolution. A promising direction is to construct *dynamic textual priors* conditioned on individual radar sequences. For example, semantic descriptions could be automatically generated or retrieved according to observed motion, intensity variation, and morphological characteristics, enabling more adaptive semantic guidance for different precipitation events.

Deeper Cross-modal Interaction. The current framework primarily integrates radar and visual representations at the feature level, while textual information serves as auxiliary guidance. Future work could explore hierarchical cross-modal interaction, where radar-native, visual, and semantic representations are progressively aligned and exchanged at multiple modeling stages. Such interaction may further exploit the complementarity among signal-level dynamics, high-level precipitation structures, and meteorological knowledge.

Uncertainty-aware and Physics-guided Forecasting. PreMM currently focuses on deterministic prediction, whereas precipitation evolution is inherently uncertain, particularly for long forecasting horizons and rapidly developing convective systems. Extending PreMM toward probabilistic or generative forecasting could model multiple plausible future trajectories and provide uncertainty

estimates. Meanwhile, incorporating physical constraints, such as advection consistency, temporal continuity, and precipitation conservation, may further improve the meteorological plausibility of predictions.

Extension to Multi-source Meteorological Observations. The current PreMM framework mainly operates on radar observations. In future work, satellite imagery, numerical weather prediction fields, surface observations, and other meteorological variables could be incorporated to provide richer environmental context. Developing efficient multimodal fusion and adaptation mechanisms while preserving the advantages of radar-native dynamics modeling represents an important step toward more comprehensive and operational precipitation nowcasting systems.

G VISUALIZATION

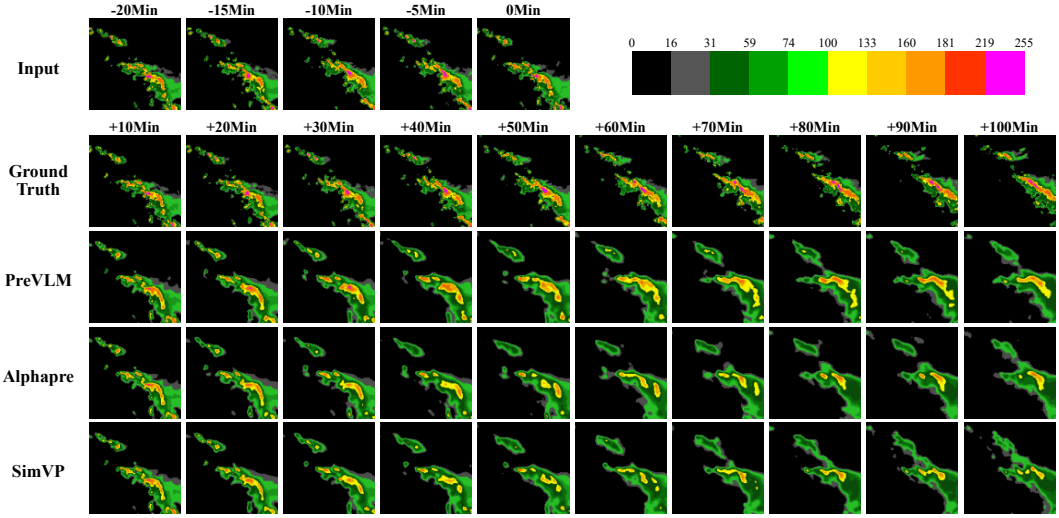

Figure 6: A visualization example from the SEVIR dataset.

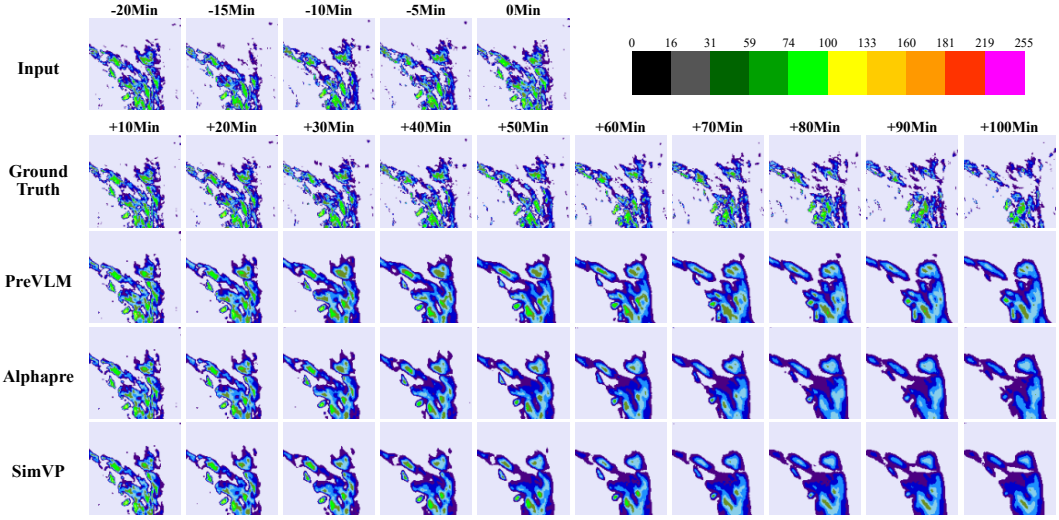

Figure 7: A visualization example from the MeteoNet dataset.

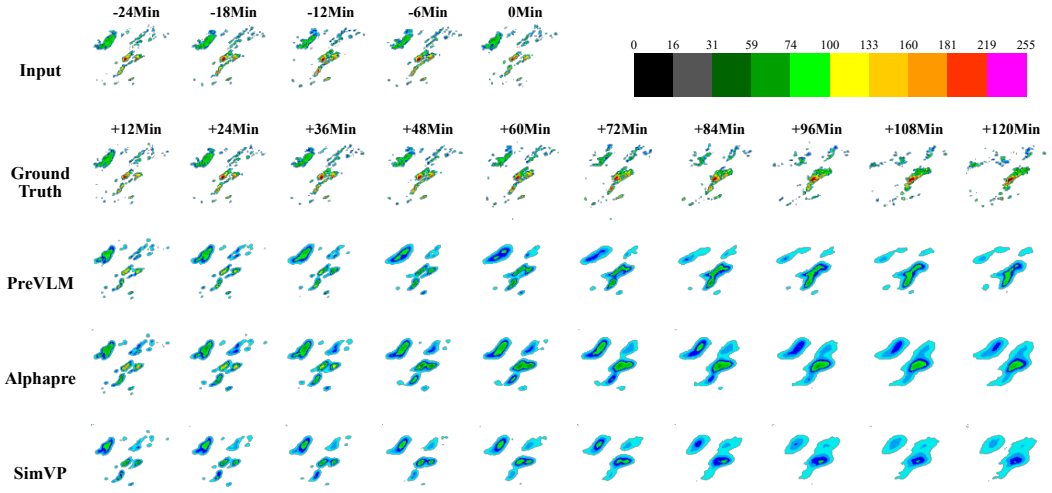

Figure 8: A visualization example from the Shanghai dataset.

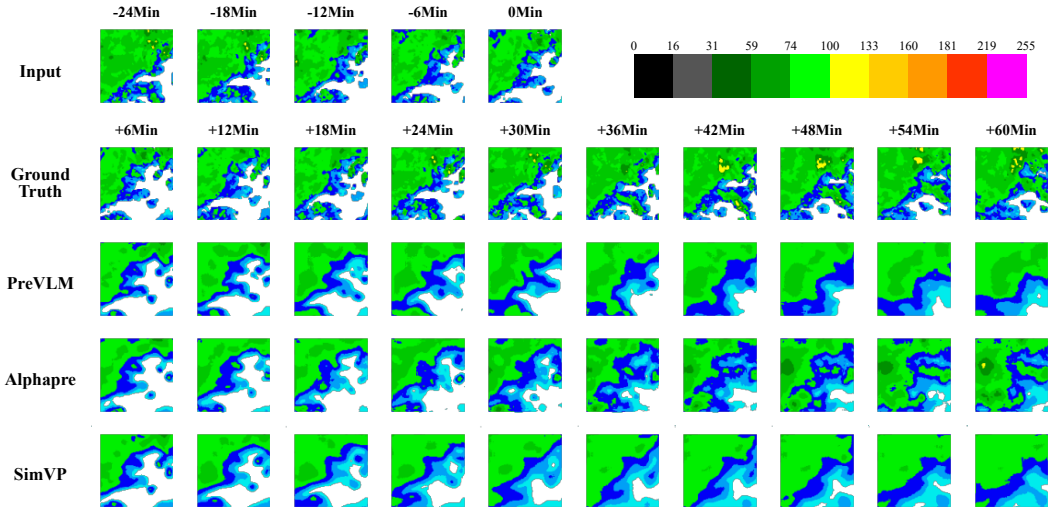

Figure 9: A visualization example from the CIKM dataset.