MIMOUDIFF: A UNIFIED MULTI-SOURCE DATA FUSION FRAMEWORK VIA MIMO UNET AND REFINED DIFFUSION FOR PRECIPITATION NOWCASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Precipitation nowcasting is a vital spatio-temporal prediction task essential for various meteorological applications, but it faces significant challenges due to the chaotic property of precipitation systems. Mainstream methods primarily rely on radar data for echo extrapolation, but over longer lead times, radar echoes mainly exhibit translation, failing to capture precipitation generation and dissipation processes. This results in blurry predictions, attenuation of high-value echoes, and positional inaccuracies issues. In the other hand, deterministic models using MSE loss often produce blurry forecasts, while probabilistic models struggle with localization accuracy. To address these challenges, we propose a multi-source data fusion framework, which integrates satellite and radar data, with former effectively complementing limitations of latter. In this framework, we leverages global motion fields to capture echo dynamics and introduces a residual diffusion mechanism to reduce memory usage by non-residual features. Various spatio-temporal models (*e.g.* RNN-based, CNN-based, and ConvRNN-based models) can seamlessly integrated into this framework. Extensive experiments on a Jiangsu dataset demonstrates significant improvements over state-of-the-art methods, particularly in short-term forecasts. *The code and models will be released.*

1 INTRODUCTION

Precipitation nowcasting has long been a challenging part of weather forecasting, focusing on providing highly localized, short-term (*e.g.*, 0-2 hours) predictions of rainfall intensity using radar echoes and other observational data Nai et al. (2024). It is crucial for a variety of applications, including issuing emergency rainfall alerts and providing weather-related guidance for agriculture and transportation Qi-liang et al. (2024). The complexity of atmospheric dynamics and associated processes complicates the accurate prediction of precipitation at both large scales and fine resolutions, making it a key area of research interest Zhang et al. (2023).

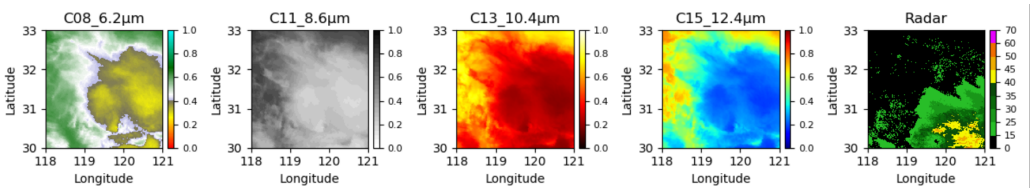

Figure 1: The temporal evolution of radar echoes and Himawari-8 satellite infrared water vapor channel C13 data, where "sat" denotes satellite and "rad" denotes radar. Mainstream models for chaotic precipitation nowcasting predominantly rely on single radar echo extrapolation. Our UniDiff method enhances forecasting by incorporating satellite data to improve predictions of strong convective development and dissipation.

Traditional numerical weather prediction (NWP) methods are computationally demanding and often impractical for very short-term forecasts due to the complexity of simulating atmospheric physical equations Tolstykh & Frolov (2005). In contrast, radar echo extrapolation methods Han et al. (2023)

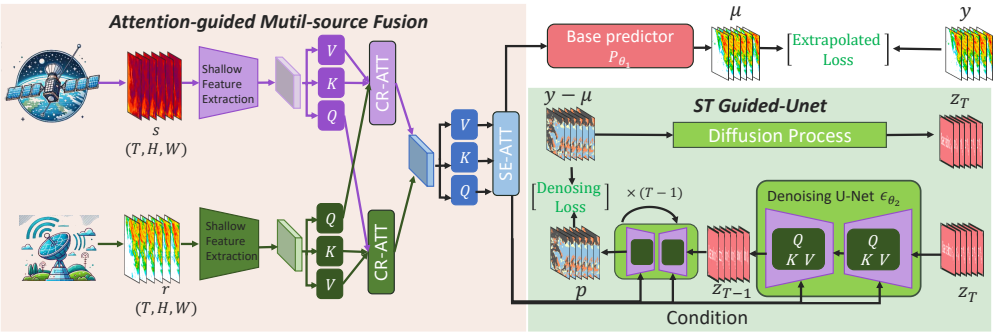

Figure 2: Illustration of our UniDiff framework for precipitation nowcasting. The framework integrates radar and satellite data through an attention-guided multi-source fusion process and applies a spatiotemporal guided UNet for diffusion-based prediction refinement. Here, s represents satellite data, and r denotes radar data. Both s and r undergo shallow feature extraction. The extracted features are then passed through a cross-attention module (CR-ATT) and a self-attention module (SE-ATT), which guide the multi-source fusion process. The fused features are further processed by the ST-Guided UNet, where a base predictor $\mathcal{P}_{\theta_1}$ generates initial coarse predictions μ. During training, the difference $y - \mu$ represents the portion that the denoising ST Guided-UNet needs to reconstruct, with the fused features serving as guiding conditions. During inference, the framework samples the residual distribution from Gaussian noise, which is then added to the coarse prediction μ to obtain the final output $\hat{y}$. This diagram represents both the training and inference processes.

offer a more computationally efficient alternative by predicting future echo patterns based on previous sequences. However, these methods typically rely on a straightforward temporal translation of radar echoes, which can result in blurring, attenuation of high-intensity echoes, and positional inaccuracies issues, particularly in extended forecasts Leinonen et al. (2023).

Over the past few years, applying deep learning techniques to precipitation nowcasting has gained significant traction, with models such as the RNN-based PredRNN++ Wang et al. (2018), ConvRNN-based TrajGRU Shi et al. (2017), and ConvLSTM Shi et al. (2015), as well as hybrid models like Rainformer Bai et al. (2022) and AA-TransUNet Yang & Mehrkanoon (2022), showing considerable promise. Nevertheless, these models are either deterministic models or probabilistic models, and all have defects to varying degrees. Deterministic models often struggle to produce sharp predictions, while probabilistic models frequently encounter challenges in achieving localization accuracy.

More recently, advancements in diffusion models have been investigated for their potential to enhance precipitation forecasting. Models like PredDiff Blücher et al. (2022), MCVD Voleti et al. (2022), Diffcast Yu et al. (2024), and SRNDiff Ling et al. (2024) integrate diffusion mechanisms to better capture the uncertainty and stochastic nature of precipitation processes. Despite these improvements, even advanced diffusion models face challenges in balancing the trade-offs between detail preservation and forecast accuracy, particularly under complex meteorological conditions.

To address these challenges, we propose a novel, flexible, and unified end-to-end framework dubbed as **UniDiff**, specifically designed for precipitation nowcasting. UniDiff seamlessly integrates radar and satellite data via an attention-guided multi-source fusion, effectively capturing and merging the complementary strengths of diverse data. Furthermore, the combined features are processed by a spatio-temporal sequence prediction module to model the primary motion patterns of the precipitation system and perform coarse forecasting. To further enhance forecast precision, UniDiff incorporates a diffusion residual component based on the RST-Unet architecture. This component leverages both deterministic models, known for their ability to capture broad spatiotemporal dynamics, and probabilistic diffusion models, which provide fine-grained accuracy, thereby generating highly refined and accurate precipitation predictions.

In summary, our main contributions are as follows:

- We introduce a novel multi-source fusion approach that effectively integrates radar and satellite data, significantly improving the accuracy of short-term precipitation forecasts by leveraging the complementary strengths of these data sources.
- We employ an RST-Unet-based diffusion residual component to refine coarse predictions, fully utilizing the strengths of deterministic models in capturing spatiotemporal dynamics and the fine-grained precision of diffusion probabilistic models.
- We design the coarse and fine processes within UniDiff to be interactive, with multi-source fusion features and coarse prediction residuals serving as conditional cues that guide the diffusion model in generating detailed and accurate predictions. This interactive approach also helps to alleviate the computational and memory demands typically associated with diffusion models.

Section II describes the related work. Section III details the architecture of our model. Section IV presents the experimental results. Conclusions are provided in Section V.

2 Related Work

2.1 Deep Learning-Based Precipitation Nowcasting

Deep learning has garnered significant attention in the domain of weather pattern analysis, offering new possibilities for precipitation nowcasting. Initially, RNN/LSTM-based methods were extensively utilized to tackle weather forecasting challenges Salman et al. (2018). Shi et al. advanced this area by integrating convolutional operations within recurrent architectures, leading to the development of the Convolutional LSTM (ConvLSTM) model Shi et al. (2015). In this model, convolutional layers replaced fully connected layers for LSTM state transitions, capturing spatiotemporal dependencies more effectively. Subsequently, Shi et al. introduced the Trajectory GRU (TrajGRU) model, which incorporates a subnetwork to dynamically learn location-variant structures for recurrent connections, achieving superior predictive accuracy on the HKO-7 precipitation benchmark Shi et al. (2017).

Further advancements in this field include the Predictive Recurrent Neural Network (PredRNN) Wang et al. (2017) and its enhanced version, PredRNN++ Wang et al. (2018). These models introduced novel mechanisms, such as the Gradient Highway Unit and Causal LSTM, to address gradient propagation challenges, thereby improving spatiotemporal prediction performance on both synthetic and real-world datasets. Following these developments, various ConvLSTM variants, such as MIM Wang et al. (2019), PFST Luo et al. (2021), and ATMConvGRU Yu et al. (2022), have been proposed. However, these RNN-based methods continue to struggle with gradient vanishing issues and require memory-intensive computations, particularly in handling long sequences Che et al. (2022).

In recent years, research has shifted towards exploring architectures that offer more efficient training and reduced computational demands. For example, Han et al. Han et al. (2020) proposed a CNN-based method that reframed the convective storm nowcasting problem as a classification task. Building on this, a UNet-based fully convolutional network (FCN) model was introduced for precipitation nowcasting Han et al. (2021), demonstrating that even simple FCN architectures can achieve performance comparable to ConvLSTM. SmaAt-UNet Trebing et al. (2021), which integrates attention modules and depthwise-separable convolutions, further enhanced predictive accuracy on real-world datasets, such as those from the Netherlands. Additionally, non-recurrent architectures like SimVP Gao et al. (2022) and PhyDNet Guen & Thome (2020) have been explored, leveraging encoding-decoding processes to make predictions. Furthermore, hybrid models combining CNNs and Transformers, such as Rainformer Bai et al. (2022) and AA-TransUNet Yimin & Mehrkanoon (2020), have shown promise in precipitation nowcasting. Despite these advances, a common challenge persists: deterministic models tend to produce blurred predictions, particularly during high-intensity precipitation events.

2.2 Conditional Diffusion in Precipitation Nowcasting

Diffusion models have emerged as a pivotal framework in generative modeling, driven by their unique approach of progressively diffusing and reconstructing noise. The foundational concepts of

diffusion models were first introduced in Sohl-Dickstein et al. (2015), but it was the introduction of Denoising Diffusion Probabilistic Models (DDPM) Ho et al. (2020) that brought widespread attention to this field. These models function by iteratively denoising Gaussian noise to learn a target distribution, allowing for content generation conditioned on various inputs, such as labels, text, or image features. The denoising network is trained to minimize the error defined by:

$$E\left[|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}}x_0 + \sqrt{1-\bar{\alpha}}\epsilon, t, c)|^2\right], \tag{1}$$

where X_0 denotes the noise-free image, $\epsilon \sim \mathcal{N}(0,1)$ represents Gaussian noise, α is a time-dependent function, and c signifies the conditioning information. The training process is centered on accurately predicting the noise added to the system, while the reverse process reconstructs the original image X_0 from the noise distribution $\mathcal{N}(0,1)$ through iterative denoising.

As diffusion models have matured, they have proven to be invaluable tools in short-term precipitation forecasting, particularly due to their denoising capabilities, which allow for the reconstruction of precise target images or sequences from noisy inputs under various conditions. For instance, the SRNDiff model Ling et al. (2024) integrates a conditional encoder to extract features from radar images, which are subsequently processed by a denoising network to produce high-resolution precipitation forecasts. This end-to-end training approach has been shown to improve prediction accuracy, especially in scenarios involving moderate to heavy rainfall. Expanding on this, the ExtDM model Zhang et al. (2024) introduces a distribution extrapolation mechanism, predicting future frames by extending the distribution of current frame features. Although initially designed for video prediction, this methodology's emphasis on temporal consistency makes it highly applicable to precipitation forecasting.

Moreover, the PredDiff model Blücher et al. (2022) combines condition-guided diffusion with domain-specific knowledge, ensuring that the forecasts not only align with historical data but also adhere to physical principles, thereby enhancing the reliability of predictions in extreme weather conditions. The Generative Diffusion Ensemble (GDE) model Asperti et al. (2023) further showcases the potential of diffusion models in handling high levels of uncertainty in weather forecasting. By generating multiple forecast scenarios based on conditional guidance, and refining these through post-processing, GDE underscores the ability to synthesize a range of potential outcomes, leading to more accurate and consistent precipitation predictions.

3 METHODOLOGY

3.1 UNIDIFF: A COARSE-TO-FINE MULTI-SOURCE FUSION FRAMEWORK FOR PRECIPITATION NOWCASTING

The UniDiff framework is developed as a novel coarse-to-fine approach to address the challenges in short-term precipitation nowcasting by leveraging both the radar and satellite data. The framework generates initial coarse predictions, which are subsequently refined to produce highly accurate precipitation forecasts. Formally, the inputs are represented by two spatiotemporal sequences $R = \{r_1, r_2, \ldots, r_n\}$ and $S = \{s_1, s_2, \ldots, s_n\}$, consisting of n radar and satellite images, respectively, with consistent spatial and temporal resolutions (10 min and 1 km). The goal of the UniDiff is to predict a sequence of m future radar echo frames $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m\}$, formulated as:

$$\hat{Y} = f_{\text{UniDiff}}(R, S), \tag{2}$$

where f_{UniDiff} represents the proposed model that first generates coarse predictions and then progressively refines them using integrated multi-source data.

The UniDiff framework consists of three core components: (1) an Attention-guided Multi-source Fusion Module, which integrates spatiotemporal features from radar and satellite data; (2) a baseline predictor responsible for generating the initial coarse prediction; and (3) a ST-Guided UNet Diffusion Module, which refines these coarse predictions to produce the final, high-resolution outputs.

3.2 ATTENTION-GUIDED MULTI-SOURCE FUSION MODULE

The Attention-guided Multi-source Fusion Module is crucial for integrating complementary spatiotemporal features from radar echo maps and satellite infrared images, thereby improving the accuracy and reliability of precipitation nowcasting.

3.2.1 SHALLOW ENCODER.

To extract relevant features from the input sequences $R = \{r_1, r_2, \ldots, r_n\}$ and $S = \{s_1, s_2, \ldots, s_n\}$, we employ the shallow encoding process. This process produces high-dimensional feature representations F_r and F_s, defined as follows:

$$F_r = \Phi_r(R), \quad F_s = \Phi_s(S), \tag{3}$$

where F_r and $F_s \in \mathbb{R}^{H' \times W' \times T' \times d}$ are the resulting high-dimensional feature maps corresponding to radar and satellite inputs, respectively.

3.2.2 CROSS-SOURCE ATTENTION (CR-ATT) LAYER.

The Cross-source Attention (CR-ATT) Layer fuses features across all dimensions by computing queries, keys, and values through linear projections:

$$\{Q_r, K_r, V_r\} = F_r W_r, \quad \{Q_s, K_s, V_s\} = F_s W_s, \tag{4}$$

where W_r and W_s are weight matrices corresponding to radar and satellite features, respectively.

The spatial interaction between radar and satellite features is facilitated by exchanging queries and computing attention-weighted values:

$$F_r^{\text{att}} = \text{softmax}\left(\frac{Q_r K_s^\top}{\sqrt{d_k}}\right) V_s, \tag{5}$$

$$F_s^{\text{att}} = \text{softmax}\left(\frac{Q_s K_r^\top}{\sqrt{d_k}}\right) V_r, \tag{6}$$

where d_k denotes the dimensionality of the queries and keys. The fused features are concatenated to form the final cross-source fused feature map:

$$F_{\text{fused}}^{\text{cross}} = \text{Concat}\left(F_r^{\text{att}}, F_s^{\text{att}}\right). \tag{7}$$

3.2.3 INTERACTION FUSION DECODER.

Following the fusion of features, the Interaction Fusion Decoder further refines the combined features through a self-attention mechanism (SE-ATT):

$$F_{\text{fused}}^{\text{self}} = \text{softmax}\left(\frac{Q_f K_f^\top}{\sqrt{d_k}}\right) V_f, \tag{8}$$

where Q_f, K_f, V_f are derived from the concatenated feature map $F_{\text{fused}}^{\text{cross}}$.

The refined fused features are then passed to the ST-Guided UNet module to generate fine-grained predictions.

Algorithm 1 Condition DDPM Training for UniDiff

Input: Dataset of samples $q(x_0)$, total timesteps T
Output: Trained model parameters θ

1: **Initialize:** Model parameters θ
2: **while** training not converged **do**
3: Sample $x_0 \sim q(x_0)$
4: Sample timestep $t \sim \text{Uniform}(\{1, \ldots, T\})$
5: Sample noise $\epsilon \sim \mathcal{N}(0, I)$
6: Compute the noisy sample:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon$$

7: Take a gradient descent step on:

$$\nabla_\theta \|\epsilon - \epsilon_\theta(x_t, t, \text{Condition})\|^2$$

8: **end while**
9: **return** Trained model parameters θ

3.3 ST-GUIDED UNET: A MULTI-SOURCE SPATIOTEMPORAL FEATURE-GUIDED DIFFUSION MODULE

The ST-Guided UNet is the cornerstone of the UniDiff framework, responsible for refining the coarse predictions generated by the baseline predictor. This module is specifically designed to handle the complex spatiotemporal dependencies inherent in precipitation nowcasting, and it employs a diffusion process to progressively refine the predictions.

ST-Guided UNet Architecture. The ST-Guided UNet leverages a hierarchical UNet architecture enhanced with spatiotemporal attention mechanisms, allowing it to effectively capture and model the intricate spatial and temporal dependencies within the data. This architecture is particularly advantageous for tasks that require the integration of multi-source data, such as precipitation nowcasting.

The ST-Guided UNet takes the fused features $F_{\text{fused}}^{\text{self}}$ as input, processing them through a series of convolutional layers and attention mechanisms. These operations generate a refined residual prediction p, which is used to enhance the initial coarse prediction μ provided by the baseline predictor $\mathcal{P}_{\theta_1}$. The final prediction $\hat{Y}$ is obtained through the following refinement process:

$$\hat{Y} = \mu + p, \tag{9}$$

where $p = \mathcal{E}(F_{\text{fused}}^{\text{self}})$ represents the residual prediction refined by the ST-Guided UNet.

3.3.1 LOSS FUNCTION.

The UniDiff framework's overall loss function is designed to optimize both the coarse prediction and the refined prediction generated by the ST-Guided UNet. The total loss function $\mathcal{L}_{\text{UniDiff}}$ is defined as:

$$\mathcal{L}_{\text{UniDiff}} = \alpha \mathcal{L}_{\text{coarse}} + (1 - \alpha)\mathcal{L}_{\text{refine}}, \tag{10}$$

where $\mathcal{L}_{\text{coarse}}$ represents the loss associated with the coarse prediction μ, and $\mathcal{L}_{\text{refine}}$ corresponds to the loss incurred by the refined prediction $\hat{Y}$. The parameter α serves as a balance factor, adjusting the contributions of the coarse and refined predictions in the total loss, thereby ensuring that the model optimizes both components effectively during training.

Latent Diffusion in UniDiff. To further enhance the coarse predictions generated by the baseline predictor, the UniDiff model employs a latent space approach for reconstructing the residual. This process involves two main stages: first, the residual $p = y - \mu$ is computed using the coarse prediction μ and the ground truth y; second, a conditional diffusion model reconstructs this residual

within the latent space, using the multi-source fused features obtained from the attention-guided fusion module as guiding conditions.

Conditional Diffusion Process: The latent diffusion model predicts the one-step-ahead noisy latent residual z_{t-1} using the conditioned latent feature z_{cond} derived from the fused features:

$$p(z_0 : T \mid z_{\text{cond}}) = p(z_T) \prod_{t=1}^{T} p_\theta(z_{t-1} \mid z_t, z_{\text{cond}}), \tag{11}$$

where z_{cond} represents the conditioned latent feature vector derived from the multi-source fused features, and p_θ represents the diffusion model responsible for generating the residual prediction.

The training of UniDiff, including the training process for the diffusion model, follows the steps outlined in Algorithm 1. The training objective in this latent space is expressed as:

$$\mathcal{L}_{\text{refine}} = \mathbb{E}_{(x,y),t,\epsilon\sim\mathcal{N}(0,I)} \|\epsilon - \epsilon_\theta(z_t, t, z_{\text{cond}})\|^2, \tag{12}$$

where ϵ_θ represents the noise predictor within the diffusion model, and z_{cond} serves as the conditioning information derived from the fused multi-source features.

This approach ensures that the diffusion model effectively reconstructs the residual p within the latent space, thereby refining the final precipitation nowcasting output by leveraging the complementary strengths of radar and satellite data.

3.3.2 Inference Process.

The inference process follows a similar sequence to the training phase, with the primary distinction being the application of the diffusion model for prediction. Initially, the latent state z_T is sampled from a standard Gaussian distribution $\mathcal{N}(0, I)$. A series of denoising steps is then performed using the learned noise predictor ϵ_θ, which iteratively refines the residual state $\hat{p}$. The final prediction $\hat{Y}$ is obtained by combining the denoised residual with the coarse prediction μ, as described in Eq. (9).

4 Experiment

4.1 Experimental Setting

4.1.1 Dataset.

The dataset utilized in this study comprises radar and satellite data collected over three years, from June to August, during the period 2019 to 2021, totaling 9 months. Both radar and satellite data are captured at a spatial resolution of 1 km and a temporal resolution of 10 minutes, producing images with dimensions of 300×300 pixels.

For the purposes of this experiment, we leverage radar echo data in conjunction with infrared channel C13 from the Himawari-8 satellite. The dataset is divided into training and testing sets based on temporal segmentation: the training set includes data from August 2019 to August 2021 (encompassing 7 months), while the testing set is derived from data collected during June and July 2019 (a total of 2 months). Following preprocessing steps, such as denoising and interpolation, a sliding window approach is employed to segment the dataset into distinct events, with each event comprising 6 frames of radar and satellite inputs, followed by 6 frames of radar outputs.

To focus specifically on significant precipitation events, the dataset is filtered by computing the mean radar reflectivity across the 6 input frames. Events where the mean value exceeds a threshold of 1 dBZ are selected for further analysis. As a result, the training set contains 5859 precipitation events, while the testing set includes 1068 precipitation events.

The data preprocessing and event selection process is detailed in Algorithm 2, which outlines the steps involved in filtering and selecting valid precipitation events based on the criteria mentioned.

Algorithm 2 Precipitation Event Filtering

Input: Radar frames R and satellite frames S (10-minute resolution, 2019-2021), Threshold $T_{avg} = 1$ dBZ, Input frames $L_{in} = 6$, Output frames $L_{out} = 6$
Output: Set of valid precipitation events $event\_set$

$event\_set \leftarrow \{\}$ {Initialize empty set for events}
$i \leftarrow 1$ {Initialize the index}
while $i + L_{in} + L_{out} - 1 \leq \text{len}(R)$ **do**
  $is\_valid\_event \leftarrow$ True {Assume event is valid}
  for $j \leftarrow 0$ **to** $L_{in} - 1$ **do**
    if Mean$(R[i + j]) \leq T_{avg}$ **then**
      $is\_valid\_event \leftarrow$ False {Invalidate event}
      break
    end if
  end for
  if is_valid_event **then**
    $event \leftarrow (R[i : i + L_{in} - 1], S[i : i + L_{in} - 1], R[i + L_{in} : i + L_{in} + L_{out} - 1])$
    add $event$ **to** $event\_set$
  end if
  $i \leftarrow i + 3$
end while
return $event\_set$

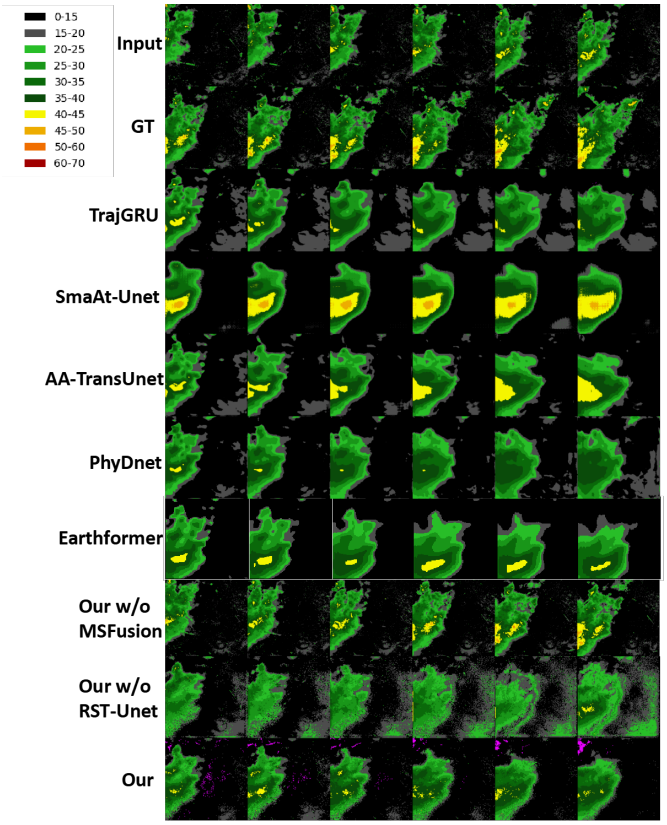

Figure 3: Qualitative comparison of predicted radar echoes between UniDiff and other SOTA models.

4.1.2 Evaluation Metrics.

The performance of our precipitation nowcasting model is evaluated using several metrics across different reflectivity thresholds (25, 35, 40, 45, 50 dBZ). Specifically, we employ the Critical Suc-

Table 1: Performance comparison of different deep learning models under **CSI** and **HSS** indexes. The best result for each metric is highlighted in bold, where "↑" imply that higher, and lower values are better.

Models	CSI↑					HSS↑				
	25dBZ	35dBZ	40dBZ	45dBZ	50dBZ	25dBZ	35dBZ	40dBZ	45dBZ	50dBZ
TrajGRU	0.588	0.406	0.283	0.184	0.139	0.690	0.545	0.421	0.301	0.238
SmaAt-Unet	0.657	0.505	0.350	0.235	0.188	0.755	0.647	0.500	0.364	0.298
AA-TransUnet	0.639	0.475	0.296	0.215	0.168	0.738	0.619	0.439	0.339	0.275
PhyDnet	0.639	0.476	0.296	0.215	0.168	0.738	0.619	0.339	0.339	0.275
Earthformer	0.649	0.450	0.167	0.027	0.001	0.754	0.605	0.280	0.052	0.002
DiffCast	0.605	0.476	0.400	0.315	0.206	0.516	0.496	0.464	0.408	0.306
Our w/o MSFusion	0.656	0.520	0.437	0.342	0.222	0.592	0.551	0.510	0.445	0.330
Our w/o RST-Unet	0.630	0.605	0.543	0.469	0.388	0.743	0.732	0.684	0.623	0.549
Our	0.529	0.433	0.335	0.263	0.174	0.652	0.580	0.479	0.395	0.276

cess Index (CSI), Heidke Skill Score (HSS), False Alarm Ratio (FAR), and Probability of Detection (POD) to assess the accuracy of precipitation predictions. Additionally, the Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) are used to evaluate the quality of the generated radar images. Together, these metrics provide a comprehensive evaluation of the model's ability to accurately forecast precipitation and generate high-quality radar imagery.

4.1.3 Training Details.

We train the UniDiff framework for 200,000 iterations using the Adam optimizer with a learning rate of 0.0001. The diffusion model follows standard configurations with 1000 diffusion steps and 250 denoising steps for inference, utilizing the DDIM Song et al. (2020) sampler. To balance the contributions of deterministic loss and denoising loss during training, we set the loss weight factor $\alpha = 0.5$. All experiments are conducted on a system equipped with a single Tesla V100 GPU.

Table 2: Performance comparison of different deep learning models under **FAR** and **Image Quality** indexes. The best result for each metric is highlighted in bold, where "↑" imply that higher, and lower values are better.

Models	FAR↓					Image Quality	
	25dBZ	35dBZ	40dBZ	45dBZ	50dBZ	SSIM ↑	LPIPS ↓
TrajGRU	0.345	0.563	0.689	0.778	0.809	0.513	0.629
SmaAt-Unet	0.230	0.398	0.509	0.610	0.661	0.563	0.527
AA-TransUnet	0.270	0.392	0.511	0.620	0.622	0.548	0.568
PhyDnet	0.269	0.392	0.511	0.620	0.708	0.548	0.102
Earthformer	0.175	0.240	0.240	0.140	0.087	0.291	0.610
DiffCast	0.222	0.286	0.339	0.397	0.501	0.180	0.272
Our w/o MSFusion	0.181	0.242	0.290	0.341	0.441	0.189	0.284
Our w/o RST-Unet	0.182	0.214	0.243	0.284	0.331	0.285	0.382
Our	0.265	0.160	0.139	0.120	0.167	0.137	0.393

4.2 Comparison with SOTAs

4.2.1 Visualization Comparison.

Figure 3 presents a qualitative comparison of the predicted radar echoes between UniDiff and several state-of-the-art (SOTA) models, including TrajGRU, SmaAt-Unet, AA-TransUNet, Phydnet, and MCVD. The visual results clearly demonstrate that UniDiff outperforms the other models in maintaining finer details and better capturing the spatiotemporal structures of precipitation events

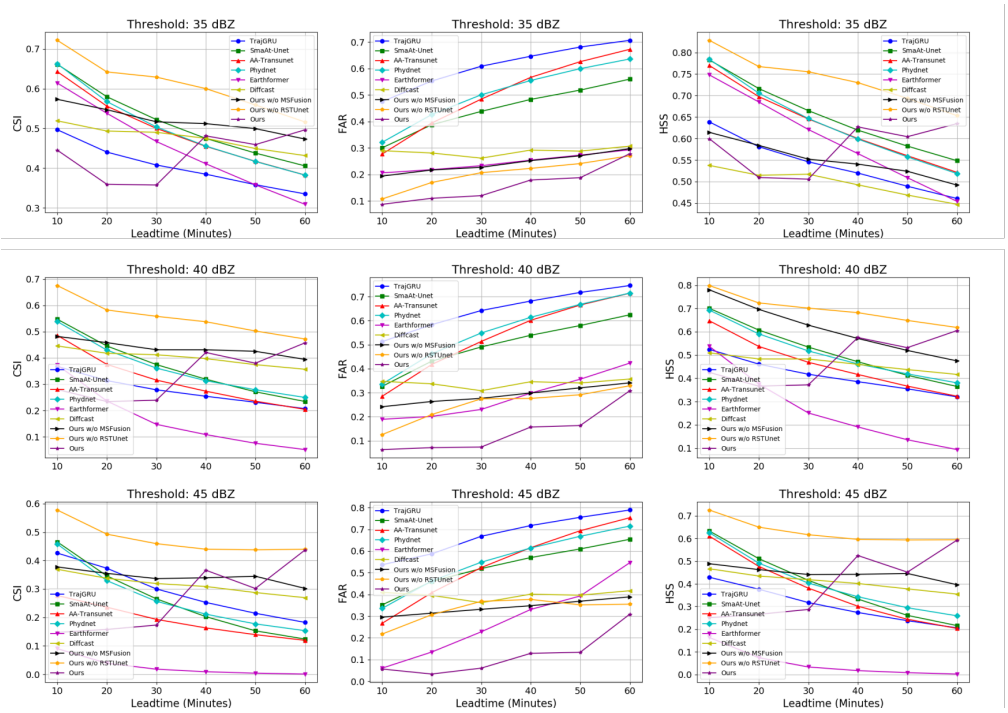

Figure 4: Comparison of CSI and HSS metrics for different SOTA models at various dBZ thresholds.

over time. The enhanced visual fidelity is a direct result of the attention-guided multi-source fusion and the diffusion-based refinement processes integrated within UniDiff, which enable more realistic and accurate precipitation nowcasting.

4.2.2 Per-frame and Threshold-based Comparison.

To further evaluate UniDiff's performance, we conducted per-frame and threshold-based comparisons, as illustrated in Figure 4 and detailed in Tables 1 and 2. UniDiff consistently achieves higher scores in Critical Success Index (CSI) and Heidke Skill Score (HSS) across various dBZ thresholds (25, 35, 40, 45, 50 dBZ) compared to the SOTA models. Notably, UniDiff demonstrates superior performance at higher dBZ thresholds, which are crucial for accurately predicting intense precipitation events. In addition to these improvements, UniDiff also exhibits a lower False Alarm Ratio (FAR) and a higher Probability of Detection (POD) across different thresholds, further validating its robustness and reliability in diverse precipitation intensity scenarios. These results underscore UniDiff's effectiveness in both temporal stability and accuracy across different lead times and thresholds, making it a reliable model for real-world forecasting tasks.

4.3 Ablation Studies

To assess the importance of individual components within the UniDiff framework, we performed ablation studies by systematically removing key modules, specifically the satellite data input (*Our* w/o *satellite*), the attention-guided multi-source fusion module (*Our* w/o *MSFusion*), and the ST-Guided UNet (*Our* w/o *RST-Unet*). The results, presented in Tables 1 and 2, reveal that the removal of any of these components significantly degrades the model's performance, particularly in CSI and HSS metrics. The absence of the *MSFusion* module, which refers to the attention-guided multi-source fusion process (replaced by simple element-wise addition of the two data sources), resulted in the most substantial drop in performance, highlighting its critical role in effectively integrating radar and satellite data to enhance prediction accuracy. These ablation studies confirm that each component within the UniDiff framework is essential for achieving the high levels of accuracy and robustness observed in our experiments.

5 CONCLUSION

In this paper, we introduced UniDiff, an innovative end-to-end framework tailored for precipitation nowcasting, which harnesses the power of multi-source data integration and diffusion-based refinement. By effectively fusing radar and satellite data through an attention-guided multi-source fusion mechanism, UniDiff capitalizes on the complementary strengths of these heterogeneous data sources to enhance the accuracy and robustness of precipitation predictions. The framework's coarse-to-fine refinement strategy, utilizing an RST-Unet-based diffusion residual component, bridges the gap between deterministic and probabilistic approaches, enabling the model to capture broad spatiotemporal dynamics while preserving fine-grained details. This interactive process ensures that the coarse predictions are progressively refined, guided by conditional cues from the multi-source fusion, resulting in superior forecasting performance. Extensive evaluations on the Jiangsu dataset have demonstrated that UniDiff outperforms existing state-of-the-art models, particularly in maintaining higher accuracy as the lead time increases.

REFERENCES

A Asperti, F Merizzi, A Paparella, G Pedrazzi, M Angelinelli, and S Colamonaco. Precipitation nowcasting with generative diffusion models. arxiv 2023. *arXiv preprint arXiv:2308.06733*, 2023.

C. Bai, F. Sun, J. Zhang, Y. Song, and S. Chen. Rainformer: Features extraction balanced network for radar-based precipitation nowcasting. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

Stefan Blücher, Johanna Vielhaben, and Nils Strodthoff. Preddiff: Explanations and interactions from conditional expectations. *Artificial Intelligence*, 312:103774, 2022.

H. Che, D. Niu, Z. Zang, Y. Cao, and X. Chen. Ed-drap: Encoder–decoder deep residual attention prediction network for radar echoes. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

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*, pp. 3170–3180, 2022.

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*, pp. 11474–11484, 2020.

Daehyeon Han, Minki Choo, Jungho Im, Yeji Shin, Juhyun Lee, and Sihun Jung. Precipitation nowcasting using ground radar data and simpler yet better video prediction deep learning. *GIScience & Remote Sensing*, 60(1):2203363, 2023.

L. Han, J. Sun, and W. Zhang. Convolutional neural network for convective storm nowcasting using 3-D Doppler weather radar data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2): 1487–1495, 2020.

L. Han, H. Liang, H. Chen, W. Zhang, and Y. Ge. Convective precipitation nowcasting using u-net model. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–8, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jussi Leinonen, Ulrich Hamann, Daniele Nerini, Urs Germann, and Gabriele Franch. Latent diffusion models for generative precipitation nowcasting with accurate uncertainty quantification. *arXiv preprint arXiv:2304.12891*, 2023.

Xudong Ling, Chaorong Li, Fengqing Qin, Peng Yang, and Yuanyuan Huang. Srndiff: Short-term rainfall nowcasting with condition diffusion model. *arXiv preprint arXiv:2402.13737*, 2024.

C. Luo, X. Li, and Y. Ye. PFST-LSTM: A SpatioTemporal LSTM model with pseudoflow prediction for precipitation nowcasting. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:843–857, 2021.

Congyi Nai, Baoxiang Pan, Xi Chen, Qiuhong Tang, Guangheng Ni, Qingyun Duan, Bo Lu, Ziniu Xiao, and Xingcai Liu. Reliable precipitation nowcasting using probabilistic diffusion models. *Environmental Research Letters*, 19(3):034039, 2024.

WU Qi-liang, WANG Xing, ZHANG Tong, MIAO Zi-shu, YE Wei-liang, and LI Hao. Diffree: Feature-conditioned diffusion model for radar echo extrapolation. 2024.

A. G. Salman, Y. Heryadi, E. Abdurahman, and W. Suparta. Single layer multi-layer long short-term memory (lstm) model with intermediate variables for weather forecasting. *Procedia Computer Science*, 135:89–98, 2018.

X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, pp. 802–810, 2015.

X. Shi, Z. Gao, L. Lausen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo. Deep Learning for Precipitation Nowcasting: A Benchmark and A New Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5617–5627, 2017.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

M. A. Tolstykh and A. V. Frolov. Some current problems in numerical weather prediction. *Izvestiya Atmospheric and Oceanic Physics*, 41:285–295, 2005.

K. Trebing, S. Tomasz, and S. Mehrkanoon. Smaat-unet: Precipitation nowcasting using a small attention-unet architecture. *Pattern Recognition Letters*, 145:178–186, 2021.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022.

Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 879–888, 2017.

Y. Wang, Z. Gao, M. Long, J. Wang, and P. S. Yu. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *In Proc. Machine Learning (ICML)*, pp. 5123–5132, 2018.

Y. Wang, J. Zhang, H. Zhu, M. Long, J. Wang, and P. S. Yu. Memory in memory: A predictive neural network for learning higher-order nonstationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9146–9154, 2019.

Yimin Yang and Siamak Mehrkanoon. Aa-transunet: Attention augmented transunet for nowcasting tasks. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 01–08. IEEE, 2022.

Y. Yimin and S. Mehrkanoon. Rainformer: Attention augmented transunet for nowcasting tasks. arXiv preprint arXiv:2202.04996, 2020.

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*, pp. 27758–27767, 2024.

T. Yu, Q. Kuang, and R. Yang. ATMConvGRU for weather forecasting. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

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 (7970):526–532, 2023.

Zhicheng Zhang, Junyao Hu, Wentao Cheng, Danda Paudel, and Jufeng Yang. Extdm: Distribution extrapolation diffusion model for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19310–19320, 2024.