remote sensing

MDPI

Article

RadarEchoMamba: A Fast, High-Fidelity Pyramidal Bidirectional Mamba Model for Radar Echo Extrapolation

Huantong Geng [1], Zhanpeng Shi [1,*], Jinzhong Min [2,3], Fangli Wu [1] and Han Zhao [1]

[1] School of Computer Science & School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China
[2] School of Atmospheric Science, Nanjing University of Information Science & Technology, Nanjing 210044, China
[3] China Meteorological Administration Tornado Key Laboratory, Foshan 528000, China
* Correspondence: 202411200008@nuist.edu.cn

Highlights

What are the main findings?

- RadarEchoMamba, an end-to-end framework built on bidirectional Mamba backbone, models long-range spatiotemporal dependencies with linear computational complexity for radar echo extrapolation.
- Multi-scale tubelet embedding and context-guided progressive decoding provide multi-resolution tokens and input-derived structural priors for high-resolution reconstruction.

What are the implications of the main findings?

- Mitigating prediction blurring, intensity attenuation and artificial echo dilation improves the physical fidelity of two-hour precipitation nowcasts for severe convective weather warning.
- The lightweight variant provides a compact and low-latency configuration for the evaluated real-time nowcasting setting.

Abstract

Radar echo extrapolation is a fundamental task in precipitation nowcasting. However, existing models based on RNNs, CNNs, and Transformers are often constrained by error accumulation, the loss of temporal information, and quadratic computational complexity. Consequently, these limitations can lead to prediction blurring and the distortion of fine-grained details in extrapolated radar images. To address these challenges, we propose RadarEchoMamba, a novel extrapolation framework built on the efficient Mamba model. RadarEchoMamba uses a pyramidal architecture to capture multi-scale features, and its core is a Bidirectional Spatiotemporal Mamba backbone that models dependencies within the observed historical input window. Furthermore, we introduce a spatiotemporal prior module driven by the input sequence to guide the reconstruction of high-resolution features during the decoding phase. This design enables the model to generate detail-rich predictions, improving upon the blurring issues prevalent in traditional extrapolation models. Experimental results on the South China and Shanghai-2020 datasets show the effectiveness of our method. Compared with strong Transformer-based baselines, the lightweight variant uses fewer parameters and achieves lower measured inference latency, while maintaining competitive prediction quality. Its main advantages are observed in visual-fidelity metrics, with improved SSIM and reduced LPIPS on the South China dataset.

Keywords: state space model; radar echo extrapolation; efficient model; Mamba

Academic Editor: John Trinder

Received: 24 May 2026
Revised: 24 June 2026
Accepted: 7 July 2026
Published: 8 July 2026

1. Introduction

Radar echo extrapolation is a core technology for the precise early warning of severe convective weather, playing a critical role in safeguarding socio-economic stability and public safety. With the deployment of next-generation radar systems, the field of nowcasting is entering a new paradigm characterized by high spatiotemporal resolution. The massive volume of high-definition data reveals fine-grained physical processes with unprecedented detail, laying a solid data foundation for higher-precision forecasting. However, this increase in resolution also imposes severe challenges on existing deep learning models. On one hand, models are required to accurately parse and extrapolate large-scale advective motions to maintain the physical consistency of macroscopic structures. On the other hand, they must generate microscopic details commensurate with high-definition inputs, such as the evolution of convective cores, to ensure high prediction fidelity. Current mainstream deep learning architectures, however, often lack designs tailored for high-resolution processing, leading to prevalent performance bottlenecks when addressing these conflicting multi-scale demands.

Data-driven methods based on deep learning have emerged as the dominant paradigm in current nowcasting research. These approaches formulate the forecasting task as a high-dimensional spatiotemporal sequence generation problem, primarily realized through architectures such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Transformers. Despite significant progress, the inherent architectural characteristics of these mainstream approaches make them ill-equipped to effectively address the challenges posed by high-resolution data. Specifically, while RNNs excel at capturing local evolution, they are prone to losing long-range macroscopic structures due to error accumulation. CNN-based methods, which typically convert temporal modeling into an image-to-image mapping problem, inherently suffer from the loss of crucial temporal dynamics. Transformers, although possessing a global receptive field, are constrained by quadratic computational complexity. This computational burden makes it difficult to generate fine-grained microscopic details when processing high-resolution inputs, leading to prevalent blurring effects in the prediction results.

Recurrent Neural Networks (RNNs), represented by ConvLSTM [1] and PredRNN [2], served as the early dominant paradigm, relying on a frame-by-frame autoregressive mechanism. While this mechanism can capture local spatiotemporal evolution in low-resolution data, its limitations are severely amplified in high-resolution scenarios. Handling massive high-resolution data requires maintaining an exceedingly large hidden state space. However, the inherent recurrent dependency of RNNs precludes parallelization, thereby restricting the scalability of model capacity. More critically, the effect of error accumulation is exacerbated in high-resolution contexts, manifesting as the rapid forgetting of high-frequency information. Consequently, in long-term forecasting, these models tend to generate blurry predictions as a shortcut to minimize the regression loss.

In response to the issues of sequential training and error accumulation in RNNs, CNN-based architectures, represented by SimVP [3], reformulate the temporal forecasting problem as an image-to-image mapping task. By compressing the temporal dimension into the channel dimension, these methods leverage convolutional feature extraction to output the entire future sequence in a single, parallelized pass. This end-to-end non-autoregressive paradigm circumvents the computational bottlenecks of frame-by-frame recurrence and reduces cumulative errors. However, this approach of collapsing dynamic temporal evolution into static channel superposition can weaken the explicit temporal structure of the

data, leading to the confusion and loss of critical spatiotemporal evolutionary information during the encoding process.

To address the issue of long-range dependencies, researchers have introduced Transformer architectures [4], which have achieved broad success in the visual domain. By leveraging the self-attention mechanism to attain a global receptive field, Transformers compute relationships across different time steps and spatial structures, showing advantages in maintaining macroscopic structural consistency. However, when applied to next-generation high-resolution meteorological data, the quadratic computational complexity ($\mathcal{O}(N^2)$, where N is the length of the input token sequence) of self-attention imposes prohibitive computational costs for long-sequence processing. Consequently, to ensure computational feasibility, mainstream Transformer-based extrapolation models often adopt downsampling strategies [5–7] to reduce the number of input tokens N. This initial information thinning at the model's entrance can blend and smooth fine-grained microscopic details. This creates a central trade-off: to handle high-resolution data at feasible cost, the model may lose part of the high-frequency information it is supposed to predict. This leaves Transformer-based models proficient at macroscopic modeling but limited in their ability to generate high-fidelity microscopic structures.

Mainstream Transformer-based extrapolation models typically adopt a single-scale 3D convolutional spatiotemporal encoding to reduce the number of tokens. While this encoding strategy lowers computational demands, it can blend and smooth fine-grained microscopic details at the earliest stage of processing. As a result, although Transformer-based models excel at macroscopic modeling, their ability to generate high-fidelity microscopic structures remains limited.

In summary, the central challenge facing existing mainstream paradigms when processing high-resolution data lies in the trade-off between computational complexity and detail preservation. State Space Models (SSMs), represented by Mamba [8], provide a linear-complexity alternative for long-sequence modeling and are therefore attractive for high-resolution nowcasting. Motivated by this property, this paper proposes an efficient end-to-end extrapolation framework named RadarEchoMamba. The framework is designed to jointly address structural extrapolation and detail reconstruction through a unified architecture, mitigating the blurring and intensity attenuation issues caused by traditional models optimizing for pixel-wise average errors.

The main contributions of this paper are summarized as follows:

(1) This paper proposes an efficient bidirectional spatiotemporal state-space backbone for macroscopic structural modeling in radar echo extrapolation. Different from Mamba-related nowcasting models that incorporate Mamba modules within recurrent or hybrid prediction frameworks, RadarEchoMamba uses stacked spatiotemporal Mamba blocks as the principal backbone of an end-to-end non-autoregressive extrapolation model. The bidirectional design is restricted to the observed historical window, where forward and time-reversed contexts are processed in parallel and then integrated through learnable fusion. By decomposing spatiotemporal modeling into independent spatial scanning and temporal evolution, this architecture captures long-range dependencies with linear complexity. This design alleviates the quadratic computational bottleneck of Transformers and supports efficient modeling of macroscopic weather-system evolution.

(2) This paper designs an end-to-end architecture for microscopic detail reconstruction. It combines encoding, temporal modeling, and decoding components in a unified framework. A Multi-Scale Tubelet Embedding module provides token representations from different spatiotemporal resolutions at the input stage. The bidirectional Mamba backbone then models their temporal evolution, while a context-guided

conditional decoder injects input-derived structural priors during progressive reconstruction. This approach guides the recovery of high-frequency textures, enabling the model to generate sharp, realistic radar echoes.

(3) This paper evaluates the balance between efficiency, fidelity, and scalability. RadarEchoMamba is an end-to-end deterministic framework with two capacity settings: a lightweight version for efficiency analysis and a base version for higher-fidelity analysis. By relying on Mamba-based deterministic modeling, the framework avoids adversarial training and iterative sampling used by GAN or diffusion-based approaches. It shows linear cost growth for high-resolution sequences (512×512 pixels, 40 time steps), indicating its potential for real-time nowcasting under the evaluated setting.

2. Related Work

This section reviews related work in radar echo extrapolation and summarizes the theoretical foundations of Mamba and its emerging applications.

2.1. Radar Echo Extrapolation Models

The pioneering work of ConvLSTM [1] established the "Encoding-Forecasting" paradigm using convolutional recurrent networks. Building upon this, TrajGRU [9] introduced a learnable connection structure to capture location-variant motions, while PredRNN [2] and PredRNN-V2 [10] enhanced long-term modeling capabilities through spatiotemporal memory units. However, these RNN-based architectures are affected by inherent recurrent limitations. Their autoregressive prediction mechanism, which recursively uses the prediction from the previous timestamp as the input for the next, can accumulate and amplify prediction errors over time. This cumulative effect is particularly pronounced in long-term extrapolation, typically manifesting as a degradation in prediction sharpness and a systematic underestimation of the intensity of severe convective systems.

To address the bottlenecks of sequential training inefficiency and error accumulation in RNNs, CNN-based non-autoregressive architectures have been introduced to nowcasting. These approaches discard frame-by-frame iterative generation and instead use CNNs to model spatiotemporal sequence prediction as a direct mapping in feature space. This naturally parallel paradigm reduces error accumulation and improves computational efficiency. Representative works such as SimVP [3] translate temporal dynamics into channel variations via Inception blocks; MBFE-UNet [11] introduces 3D convolutions in the encoding stage to preserve temporal independence; and SCECA-Net [12] incorporates attention mechanisms to enhance the capture of key spatiotemporal features. Nevertheless, because these models are constrained by the "image-to-image mapping" paradigm, they tend to collapse dynamic temporal evolution into static feature-channel superposition. This approach weakens explicit temporal modeling, making it difficult to capture the intricate non-linear dynamics of weather systems.

To address the aforementioned issues, the Transformer architecture [4] based on the attention mechanism has been introduced to the field of nowcasting. Relying on its global receptive field and parallel computing capabilities, this architecture shows advantages in capturing long-range spatiotemporal dependencies. PredFormer [5] and Earthformer [6] utilize spatiotemporal data patches as basic processing units, directly modeling the relationship between any two spatiotemporal points in the sequence through the multi-head self-attention mechanism, thereby improving prediction structural consistency. Nevertheless, the quadratic computational complexity of the self-attention mechanism imposes large computational and memory bottlenecks when processing high-resolution, long-sequence data. To alleviate this issue, Swin Transformer [13] and its variants have been introduced to

reduce computational complexity by computing attention within local windows, but this also limits the model's ability to obtain an immediate global receptive field. Therefore, although Transformer-based models have mitigated the problem of error accumulation, existing methods still struggle to achieve an ideal balance across the three critical dimensions of modeling accuracy, computational efficiency, and scalability for high-resolution data. This restricts their application in operational nowcasting scenarios that have rigorous requirements for real-time performance.

2.2. State Space Model

State Space Models (SSMs) provide an efficient framework for long-sequence modeling. S4 [14] introduced structured state matrices for near-linear training and efficient recurrent inference, while Mamba [8] further added an input-dependent selective mechanism and hardware-aware parallel scanning [15]. Recent studies have extended Mamba-style modeling to visual and spatiotemporal tasks. VMamba [16] adapts selective scanning to two-dimensional visual representation, ChangeMamba [17] and SwinMamba [18] apply Mamba-based encoders to remote sensing vision tasks, and VMRNN [19] integrates Vision Mamba blocks with recurrent units for spatiotemporal forecasting. These studies show the potential of selective state-space models for efficient dependency modeling. However, existing Mamba-related nowcasting models mainly emphasize recurrent prediction or hybrid CNN–Mamba feature extraction. The use of stacked spatiotemporal Mamba blocks as the principal non-autoregressive backbone for high-resolution radar extrapolation, together with multi-scale embedding and conditional decoding, remains less explored.

3. Methodology

3.1. Preliminaries

Standard Mamba architectures excel at 1D sequence modeling with linear complexity, but their causal scanning nature is ill-suited for non-causal 2D image data. Flattening images directly disrupts critical spatial adjacencies. To bridge this gap, the Visual State Space (VSS) block incorporates a 2D Cross-Scan Mechanism (SS2D) [16], which serves as the fundamental building block of our backbone.

The SS2D mechanism transforms the input feature map $Z_{in} \in \mathbb{R}^{H\times W\times C}$ into four distinct 1D sequences by scanning along four corners: top-left, bottom-right, top-right, and bottom-left. Let $\mathcal{S}_i(\cdot)$ denote the scanning operation for the ith direction, and $\mathrm{SSM}_i(\cdot)$ represent the selective scan module with independent parameters. The process can be formulated as:

$$\begin{aligned} z^{(i)} &= \mathcal{S}_i(Z_{in}), \quad i \in \{1,2,3,4\} \\ h_t^{(i)} &= \overline{A}_i h_{t-1}^{(i)} + \overline{B}_i z_t^{(i)} \\ y_t^{(i)} &= C_i h_t^{(i)} \end{aligned} \tag{1}$$

where $z^{(i)}$ is the flattened sequence for direction i. The core parameters $(\overline{A}, \overline{B})$ are discretized from continuous system parameters using the zero-order hold method and are input-dependent to achieve selective information filtering.

Finally, the processed sequences from four directions are reverse-scanned $(\mathcal{S}_i^{-1})$ to restore the 2D spatial structure and merged to produce the output feature map Z_{out}:

$$Z_{out} = \sum_{i=1}^{4} \mathcal{S}_i^{-1}(y^{(i)}) \tag{2}$$

The inverse scanning operation $(\mathcal{S}_i^{-1})$ effectively reverses the spatial rearrangement (e.g., flipping or transposing) performed in the scanning step, ensuring that features from

all directions are spatially aligned before summation. This design allows any pixel in the radar image to integrate contextual information from all other pixels in four sweeping directions, establishing a global receptive field with $\mathcal{O}(N)$ complexity, where $N = H \times W$.

3.2. Overall Architecture

To address prediction blurring and microscopic detail loss in long-term extrapolation, this paper proposes an end-to-end forecasting framework named RadarEchoMamba. As illustrated in Figure 1, the framework is organized around three complementary components rather than a single reconstruction module. Mathematically, the model aims to map the input historical radar sequence $X_{in} \in \mathbb{R}^{B \times T_{in} \times C \times H \times W}$ to a high-fidelity prediction sequence $Y_{pred} \in \mathbb{R}^{B \times T_{out} \times C \times H \times W}$.

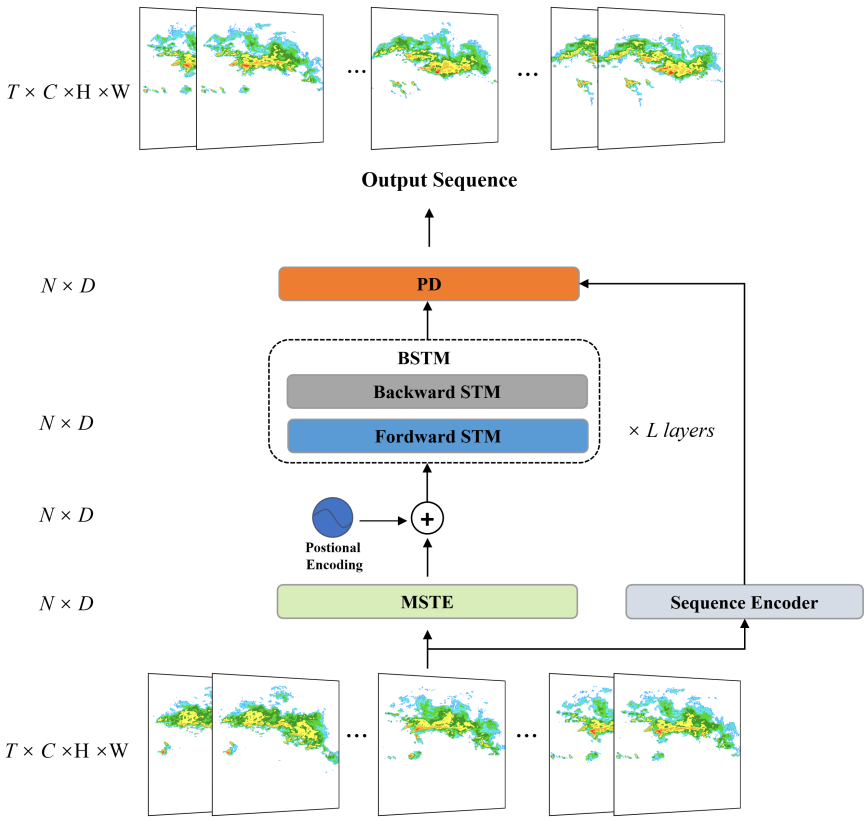

Figure 1. Overall architecture of the proposed RadarEchoMamba. It consists of three core components: the Multi-Scale Tubelet Embedding (MSTE) for feature extraction, the Bidirectional Spatiotemporal Mamba (BSTM) backbone for global modeling, and the Progressive Decoder (PD) for high-fidelity reconstruction. Additionally, a Sequence Encoder is introduced to inject structural priors into the decoder. $\oplus$ denotes element-wise summation. T, C, H, and W represent the number of time steps, channels, height, and width. N and D denote the token sequence length and embedding dimension.

The input historical radar sequence X_{in} is first processed by the Multi-Scale Tubelet Embedding (MSTE) module. MSTE extracts features at different spatiotemporal scales in parallel, so that the subsequent backbone receives token representations from multiple temporal and spatial resolutions. The input data are thereby transformed into a sequence of embedded tokens $E_{in} \in \mathbb{R}^{B \times N \times D}$.

The embedded sequence is then augmented with absolute positional encodings and fed into the Bidirectional Spatiotemporal Mamba Backbone. This backbone consists of stacked SpaceTimeMambaBlocks and serves as the principal temporal modeling module of RadarEchoMamba. It scans the observed historical sequence in both forward and backward

directions, producing a latent representation Z_{pred} that summarizes temporal evidence from both directions within the input window.

Finally, Z_{pred} is forwarded to a Progressive Conditional Decoder for predictive reconstruction. The design of this decoder is inspired by diffusion models [20,21] and aims to overcome the inherent limitations of traditional decoders in recovering high-frequency details. Unlike conventional decoders that rely only on the final latent representation, the proposed decoder additionally accepts multi-scale feature maps extracted from the original input sequence X_{in} as spatial structural priors. These priors are derived via temporal average pooling to reduce short-term dynamic variations and retain the stable macroscopic morphology over the input period. In this design, the Mamba backbone is responsible for temporal dependency modeling, while the decoder uses input-derived structural references during progressive reconstruction.

The complete pipeline can be summarized as follows. The MSTE module first tokenizes the input sequence into a compact embedding. Absolute positional encodings are then added to inject spatiotemporal location information. In parallel, the SequenceEncoder extracts multi-scale structural priors $\{\mathcal{C}_{feat}^{(k)}\}_{k=1}^{M}$ directly from X_{in} via per-frame convolutions and temporal averaging. The decoder then fuses both streams to produce the final prediction. The information flow of the framework can be formulated as:

$$\hat{E}_{\text{in}} = MSTE(X_{in}) + PE \tag{3}$$

$$\mathcal{C}_{\text{feat}} = \text{SeqEncoder}(X_{in}) \tag{4}$$

$$Z_{\text{pred}} = Backbone(\hat{E}_{\text{in}}) \tag{5}$$

$$Y_{\text{pred}} = \text{Decoder}\left(Z_{\text{pred}},\ \left\{\mathcal{C}_{\text{feat}}^{(k)}\right\}_{k=1}^{M}\right) \tag{6}$$

where $\hat{E}_{\text{in}} \in \mathbb{R}^{B\times N\times D}$, $\mathcal{C}_{\text{feat}}^{(k)} \in \mathbb{R}^{B\times C_k\times H_k\times W_k}$, and $Z_{pred} \in \mathbb{R}^{B\times N\times D}$.

3.3. Multi-Scale Tubelet Embedding Module

Traditional spatiotemporal data embedding methods typically employ a single, fixed sampling scale, which may lose critical microscopic details or macroscopic information when processing high-resolution, long-duration radar echoes. Inspired by VideoGPT [22] and VideoMamba [23], this paper designs the Multi-Scale Tubelet Embedding (MSTE) module. As shown in Figure 2, this module uses three parallel, non-weight-sharing 3D convolutional branches to extract local details, regional dynamics, and global contexts at the initial encoding stage. This design ensures that the subsequent backbone network receives a hierarchically rich initial feature field, which supports the modeling of complex spatiotemporal evolution.

Rather than relying on purely empirical settings, the scale partitioning in the MSTE module is grounded in a physically guided framework based on the spatiotemporal evolution characteristics of meteorological radar echoes. After downsampling the original radar grids to 512 × 512, the input data have a spatial resolution of 600 m per pixel and a temporal resolution of 6 min (South China dataset). Consequently, the three branches employ 3D convolutions with both kernel size and stride equal to k_s, k_m, and k_l, which are mapped to classic meteorological mesoscale definitions:

Fine-Grained Spatiotemporal Detail Branch $k_s = [1, 16, 16]$: This branch operates at a physical receptive field of 9.6 km ×9.6 km spatially and 6 min temporally. It falls within the meso-γ scale (2–20 km), focusing on rapid, highly localized physical processes such as the life cycles of individual convective cores, rapid generation and dissipation of intense echoes, and fine internal textures.

Medium-Scale Structure Branch $k_m = [2, 32, 32]$: Operating at a physical scale of 19.2 km $\times$ 19.2 km and 12 min, this branch aligns with the upper limit of the meso-γ scale. It serves as the baseline resolution, responsible for extracting multi-cell convective clusters, regional structural morphologies, and the boundaries transitioning into weak echo edges.

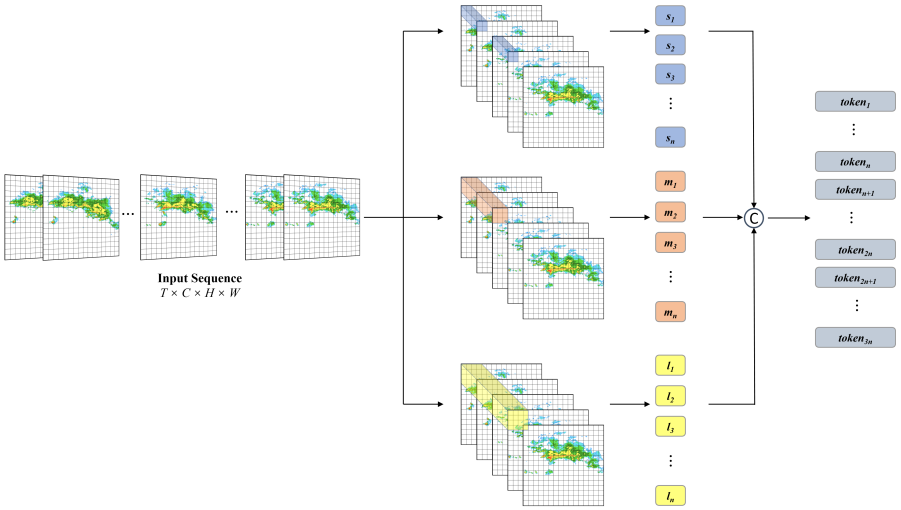

Figure 2. Schematic of the MSTE module. The input sequence is fed into three parallel branches with distinct tubelet sizes (small, medium, and large) to extract multi-scale spatiotemporal features. These features are subsequently concatenated and mapped into a unified token sequence. © represents the concatenation operation.

Macroscopic Context Branch $k_l = [4, 64, 64]$: Expanding the receptive field to 38.4 km $\times$ 38.4 km and 24 min, this branch shifts into the meso-β scale (20–200 km). It has the longest temporal duration and largest spatial receptive field, providing macroscopic contextual information, stable stratiform precipitation regions, and the broader advection trends of mesoscale convective systems.

The output of each branch first undergoes non-linear activation and GroupNorm to enhance feature expression and stability. Subsequently, to fuse these feature maps of different resolutions, the outputs of the Fine-Grained and Macroscopic branches are adjusted to the same spatiotemporal resolution as the Medium-Scale branch via nearest-neighbor temporal interpolation and bilinear spatial interpolation. Finally, the three size-aligned feature maps are concatenated along the channel dimension and deeply fused through a convolutional module to form a unified, information-rich feature map. This feature map is ultimately rearranged into the token sequence $E_{in} \in \mathbb{R}^{B \times N \times D}$, which serves as the input for the subsequent backbone network.

3.4. Bidirectional Spatiotemporal Mamba Backbone

As shown in Figure 3, the core of the RadarEchoMamba model is a bidirectional backbone network formed by the deep stacking of L spatiotemporal Mamba modules. To construct a more complete spatiotemporal context within the observed sequence, this network processes forward and backward spatiotemporal sequences in parallel. Its core state-space mechanism compresses directional sequence history into a hidden state. By fusing the outputs of the forward and backward branches, the model integrates contextual information propagated from both temporal directions, thereby obtaining a broad receptive field over the historical window.

Unlike models that employ self-attention, the framework in this paper is based on the State Space Model (SSM), whose core principle is to compress and transmit historical information through a hidden state h. In standard SSMs such as Mamba, the state at the

current time step h_t is determined solely by the state at the previous time step h_{t-1} and the current input x_t.

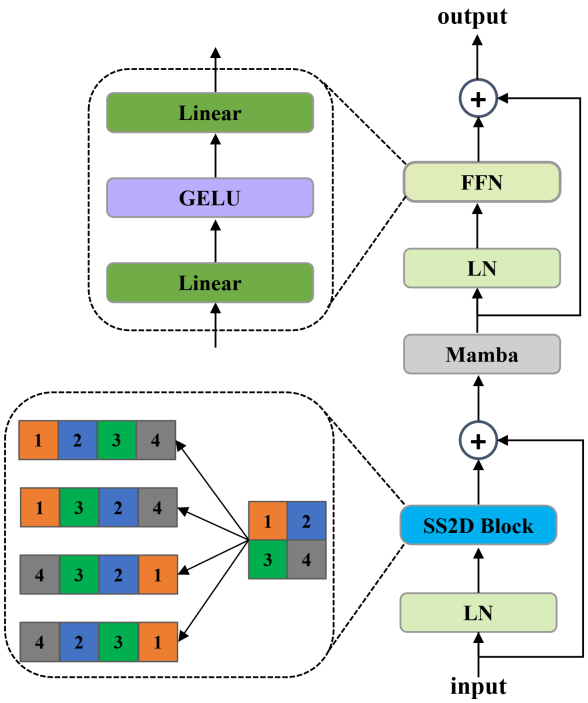

Figure 3. Architecture of the Spatiotemporal Mamba Block. The block adopts a serial spatiotemporal separation design, sequentially processing features through the Spatial Modeling Layer (SS2D), the Temporal Modeling Layer (Mamba), and the Feature Transformation Layer (FFN).

The basic building blocks for both backbones are the SpaceTimeMambaBlock, which adopts a hybrid, serial spatiotemporal separation processing strategy. Given that the temporal evolutionary laws of radar echoes (such as movement and intensity changes) are physically distinct from their spatial patterns (such as morphological structure and texture), this module decomposes the complex spatiotemporal modeling task into three sequentially executed sub-layers:

(1) Spatial Modeling Layer: To precisely capture the intricate spatial dependencies within each frame, the input token sequence $E_{in} \in \mathbb{R}^{B\times N\times D}$ is first reshaped into frame-wise batches $T_s \in \mathbb{R}^{(B\cdot N_t)\times N_h\times N_w\times D}$ and processed using the SS2D [16] module, which is specifically designed for 2D vision tasks. As formulated in Equations (1) and (2), the key advantage of SS2D lies in its unique four-direction scanning mechanism, enabling efficient aggregation of anisotropic spatial contexts. This makes it particularly effective in capturing the irregular morphologies and fine textures present in radar echoes.

(2) Temporal Modeling Layer: After spatial information fusion, the features are rearranged into spatial-point-wise batches $T_t \in \mathbb{R}^{(B\cdot N_s)\times N_t\times D}$, where $N_s = N_h \times N_w$. This operation treats the data as N_s independent time series, each representing the evolutionary history of a spatial location. The model employs standard 1D Mamba modules to process these sequences. Unlike RNNs that rely on fixed recurrent structures, the Selective SSM mechanism of 1D Mamba can dynamically adjust its state transitions based on the input content. This provides greater flexibility when modeling non-linear intensity mutations and long-term evolutionary trends in radar echoes.

(3) Feature Transformation Layer: After completing the information interaction in the two orthogonal dimensions (spatial and temporal), the data enters a standard Feed-Forward Network (FFN). The FFN independently performs a non-linear decoupling and remapping on the feature vector of each spatiotemporal token. The introduction of the FFN provides the model with the capability to perform deep feature transformations at each spatiotemporal point, enhancing the model's non-linear expressive power.

In the implementation of each sub-layer, the model follows the design paradigm of Pre-Layer Normalization and Residual Connections, which is commonly used in Vision Transformer [24] and its subsequent works, to support gradient stability and training efficiency throughout the deep network.

Let the token sequence from the MSTE module be $E_{in} \in \mathbb{R}^{B\times N\times D}$. The processing flow of the bidirectional backbone is formulated as follows:

1. Forward Path: The forward backbone, denoted as $\mathcal{F}_{fwd}$, processes the token sequence in its natural order to capture causal dependencies:

$$Z_{fwd} = \mathcal{F}_{fwd}(E_{in}; \theta_{fwd}) \tag{7}$$

where θ_{fwd} represents the parameters of the forward path, and $Z_{fwd} \in \mathbb{R}^{B\times N\times D}$ is the resulting forward latent representation.

2. Backward Path: To model anti-causal dependencies, the backward path operates on a time-reversed version of the same input sequence E_{in}. Specifically, the input sequence is first reshaped to expose the temporal dimension ($N = N_t \times N_s$), flipped along the time axis, and then reshaped back to the original token sequence format:

$$E_{rev} = \text{Reshape}(\text{Flip}(\text{Reshape}(E_{in}))) \tag{8}$$

where the reversed sequence $E_{rev} \in \mathbb{R}^{B\times N\times D}$ is then processed by the independent backward backbone $\mathcal{F}_{bwd}$:

$$Z_{bwd} = \mathcal{F}_{bwd}(E_{rev}; \theta_{bwd}) \tag{9}$$

To align the output, the resulting backward latent representation Z_{bwd} undergoes the inverse transformation (reshape, flip, and reshape back):

$$Z_{align} = \text{Reshape}(\text{Flip}(\text{Reshape}(Z_{bwd}))) \tag{10}$$

3. Fusion: Finally, the forward representation Z_{fwd} and the aligned backward representation Z_{align} are concatenated along the feature dimension and fused via a linear layer $\mathcal{W}_{fusion}$ to produce the final, holistic latent representation Z_{pred}:

$$Z_{pred} = \mathcal{W}_{fusion}\Big(\text{Concat}(Z_{fwd}, Z_{align})\Big) \tag{11}$$

Z_{fwd} mainly summarizes information propagated from the earlier part of the observed sequence. In contrast, Z_{align} summarizes information propagated from the later part after time reversal and alignment. The concatenation operation keeps these two directional summaries separated before fusion, instead of forcing them into a fixed average. The subsequent linear layer can then learn different weights for forward and backward contexts. Through this mechanism, each token in Z_{pred} integrates complementary temporal evidence from the observed historical window, providing a more complete basis for the subsequent decoding stage.

3.5. Conditional Decoder

Conventional spatiotemporal prediction models, whether based on convolutional recurrent networks or local attention mechanisms, typically employ decoders designed as independent reconstruction modules. These decoders rely primarily on the highly abstract latent representation Z_{pred} generated by the backbone to synthesize future frames. While such mechanisms can capture basic dynamic trends, they may be limited when attempting to recover high-frequency spatial details, such as fine-grained echo textures or precise morphological structures. This shortfall arises from the lack of a global, high-level spatial structural reference to guide the generation, making the predictions prone to blurring effects associated with information decay and mean-reversion during optimization.

To mitigate this issue, this paper introduces a Progressive Conditional Decoder. As shown in Figure 4, the core objective is to use multi-scale information from the original input sequence X_{in} as structural priors during reconstruction. The design comprises two key phases: Multi-scale Context Encoding and Hierarchical Feature Injection.

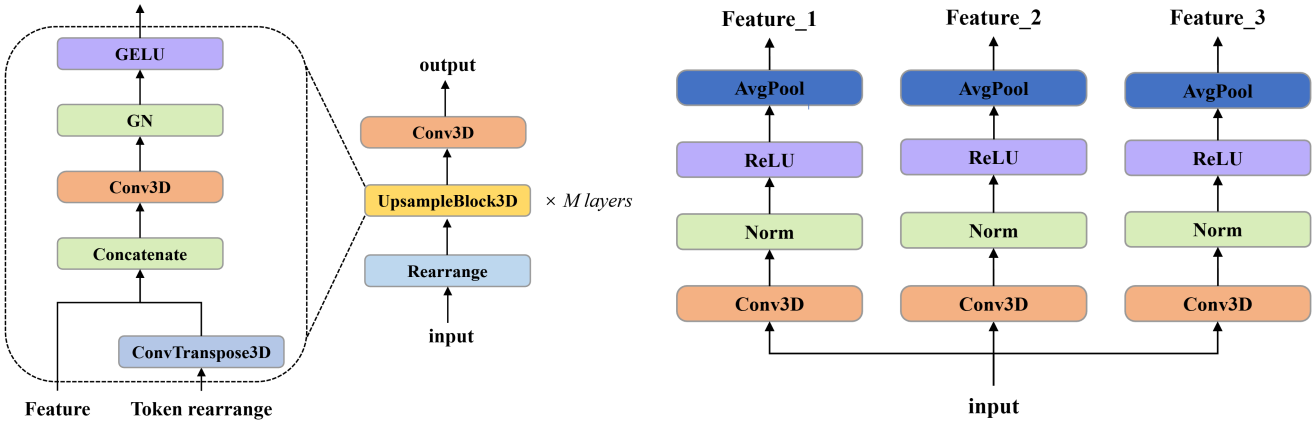

Figure 4. Details of the Conditional Decoder. (**a**) Progressive Decoder reconstructs the high-resolution prediction through M stacked upsampling layers. (**b**) Sequence Encoder extracts hierarchical spatial priors from the original input sequence.

Multi-scale Context Encoding: The decoder first distills a set of stable structural priors from X_{in}. The SequenceEncoder is not intended to preserve all temporal variations in the observed sequence. Instead, it provides a static macroscopic spatial blueprint for the decoder. Within the short input window, the macroscopic morphology, orientation, and broad intensity distribution of radar echoes usually evolve more smoothly than frame-level fluctuations. Temporal average pooling is therefore used after per-frame convolutional encoding. It acts as a temporal low-pass filter that suppresses short-term jitter and transient intensity variations while retaining the dominant spatial structure of the input period. In this division of labor, the SequenceEncoder supplies input-derived structural references, whereas the bidirectional Mamba backbone models temporal evolution and dynamic changes. The SequenceEncoder processes the input sequence through a series of convolutions and progressive downsampling to generate pyramidal feature maps at multiple scales. Let $\mathcal{F}_{seq}^{(k)}(\cdot)$ denote the convolutional transformation of the SequenceEncoder at the kth scale. The spatial prior $\mathcal{C}_{feat}^{(k)}$ is derived as follows:

$$\mathcal{C}_{feat}^{(k)} = \frac{1}{T_{in}} \sum_{t=1}^{T_{in}} (ReLU(Norm(\mathcal{F}_{seq}^{(k)}(X_{in})))) \tag{12}$$

The resulting $\mathcal{C}_{feat}^{(k)}$ is a time-averaged structural representation rather than a complete encoding of the input dynamics. It is therefore used as a spatial guide for reconstruction, while temporal changes are represented by Z_{pred} from the backbone.

Hierarchical Feature Injection: In this phase, the decoder fuses the extracted $\mathcal{C}_{feat}$ with the backbone's output Z_{pred} via a progressive upsampling scheme. Unlike conventional decoders that rely on a single information stream, our decoder operates as a dual-stream input system. At each level of the decoding path, rather than merely upsampling the features, the resolution is first increased to the target scale, and then the corresponding static structural priors from the SequenceEncoder are injected via concatenation. Let F_k be the output feature at the kth decoding stage and $\mathrm{UP}(\cdot)$ denote the spatial upsampling operation. The feature fusion process is defined as:

$$\mathbf{F}_k = \mathrm{Block}_k\left(\mathrm{Concat}[\mathrm{UP}(\mathbf{F}_{k-1}), \mathcal{C}_{feat}^{(k)}]\right) \tag{13}$$

where Block_k consists of convolutional, normalization, and activation layers that adaptively aggregate the evolutionary information from the backbone with the structural guidance from the input sequence.

This feature-injection strategy allows the decoder to receive spatial structural constraints at every resolution level. Low-resolution priors support macroscopic morphological accuracy, while high-resolution priors provide references for fine-texture recovery. By using concatenation rather than simple addition, the model can learn adaptive weights between the static priors and the dynamic features through convolutional kernels. This helps ground the reconstruction process in the observed input sequence while retaining the temporal-evolution information produced by the backbone.

4. Experiments

4.1. Dataset

The dataset used in this study is derived from observations collected by multiple Sband weather radars in the South China region between 2021 and 2023 and Shanghai-2020 dataset [25]. The composite reflectivity factor product was selected; after multi-radar mosaicking, it was processed into grid data with a spatial resolution of 300 m and a temporal resolution of 6 min, with each sample covering a range of 1024 × 1024 pixels.

To ensure the high quality and representativeness of the dataset, a rigorous data filtering process was implemented. First, combined with precipitation records from automatic weather stations, invalid or weak precipitation events with echo coverage areas below a specific threshold were initially removed. Second, to guarantee that each sample contains a complete and meaningful evolutionary process, radar sequences that are temporally continuous and have a length of no less than 40 time steps were further selected. Finally, all machine-filtered sequences underwent manual review to eliminate any anomalous data that did not meet the requirements.

Following the aforementioned processing, the final dataset consists of three years of independent sequences. To avoid data leakage and assess the model's seasonal generalization capability, the test set was independently constructed by randomly sampling from March to September of each year, with the remainder serving as the training and validation set. In all experiments, data input to the model was downsampled from 1024 × 1024 to 512 × 512 resolution via bilinear interpolation to balance computational efficiency and model performance.

The Shanghai-2020 dataset initially comprises 43,000 precipitation events, with each event consisting of 20 consecutive radar frames. Following the same data cleaning strategy applied to the South China dataset, the filtered sequences were partitioned into subsets

for model development. Specifically, 7715 sequences were allocated for training, 1803 for validation, and 1390 for testing. For experimental consistency, all radar frames were resized to a resolution of 512×512 pixels.

4.2. Experimental Setup

4.2.1. Implementation Details

The RadarEchoMamba model proposed in this paper adopts specific parameter configurations in its core modules. In the Multi-Scale Tubelet Embedding (MSTE) module, the kernel size and stride of the three parallel 3D convolutional branches are set to $[1, 16, 16]$ (small scale), $[2, 32, 32]$ (medium scale), and $[4, 64, 64]$ (large scale), respectively, to achieve parallel extraction of spatiotemporal features at different scales.

For the backbone network, this paper adheres to a flexible architectural design philosophy informed by recent advances in efficient deep learning [26]. While traditional scaling laws [27] suggest that model performance correlates positively with total parameter count, deploying such models for operational nowcasting requires a careful trade-off between modeling capacity and inference speed.

To investigate these trade-offs systematically, this paper establishes two distinct scaling dimensions. First, to prioritize operational efficiency, we design RadarEchoMamba-Light with a compact embedding dimension of 256 and a depth of 8 blocks, serving as a lightweight baseline for resource-constrained environments. Second, to explore the upper limits of capturing complex non-linear weather dynamics, we expand the embedding dimension to 768 to form RadarEchoMamba-Base. Following the deep and narrow principle in [26], we further scale the model capacity by increasing the network depth rather than width. Consequently, RadarEchoMamba-M and RadarEchoMamba-L are constructed by stacking 12 and 16 bidirectional spatio-temporal Mamba blocks, respectively, on top of the Base configuration.

4.2.2. Experimental Environment

All experiments were conducted on an Ubuntu operating system based on the PyTorch 2.7.1 framework, using a single NVIDIA RTX 4090 GPU for training and evaluation. The AdamW optimizer [28] was adopted to optimize model parameters, with the initial learning rate set to 1×10^{-4} and the weight decay coefficient set to 1×10^{-2}. To ensure training stability, a cosine annealing learning rate schedule with 10% warmup was employed [29]. The training batch size for all models was set to 1, and they were uniformly trained for 50 epochs. Regarding the loss function, the standard Mean Squared Error (MSE) Loss was used to calculate the discrepancy between the predicted values and the ground truth.

4.2.3. Evaluation Metrics

To comprehensively evaluate the performance of RadarEchoMamba, this paper employs two distinct categories of metrics: threshold-based meteorological accuracy metrics and perception-based image quality metrics.

Following standard protocols in radar nowcasting, we first binarize the radar reflectivity maps using intensity thresholds of 10, 20, and 30 dBZ. By comparing the predicted masks with the ground truth, we determine the numbers of true positives (TP), false negatives (FN), and false positives (FP). Following the metrics proposed by [30], we use the critical success index (CSI), probability of detection (POD), and false alarm rate (FAR). CSI provides a comprehensive measure of the overlap between the forecast and observation, where a higher value indicates better overall accuracy. POD quantifies the model's ability to capture actual events and represents the recall rate of the forecast. FAR reflects the degree of false positives. The structural similarity index (SSIM) [31] evaluates image similarity based on luminance, contrast, and structure. Learned Perceptual Image Patch Similarity (LPIPS) [32]

calculates the distance between deep features extracted by a pre-trained convolutional network. It aligns more closely with human visual perception of blurriness and distortion and is used to assess high-fidelity reconstruction.

$$\begin{aligned} CSI &= \frac{TP}{TP+FN+FP} \\ POD &= \frac{TP}{TP+FN} \\ FAR &= \frac{FP}{TP+FP} \\ SSIM(\hat{Y},Y) &= \frac{(2\mu_{\hat{Y}}\mu_Y + C_1)(2\sigma_{\hat{Y}Y} + C_2)}{(\mu_{\hat{Y}}^2 + \mu_Y^2 + C_1)(\sigma_{\hat{Y}}^2 + \sigma_Y^2 + C_2)} \end{aligned} \tag{14}$$

where $\mu_{\hat{Y}}$ and μ_Y represent the mean intensities of the prediction and ground truth; $\sigma_{\hat{Y}}$ and σ_Y denote their standard deviations; and $\sigma_{\hat{Y}Y}$ is the covariance between them. C_1 and C_2 are stability constants. In this paper, torchmetrics is used to calculate SSIM.

To complement the separate evaluation of detection accuracy and perceptual fidelity, we further report FCSI (Fidelity-aware Critical Success Index) as an auxiliary composite statistic. FCSI is not intended to replace CSI, POD, FAR, SSIM, or LPIPS; instead, it provides a compact summary of the balance between threshold-based overlap accuracy and image fidelity. The motivation is that a model may obtain a high CSI by expanding or smoothing the predicted echo region, while still losing structural detail and perceptual fidelity in the continuous radar image. We first define a FI (Fidelity Index) as an exponential penalty over the joint error of both perceptual metrics:

$$FI = e^{-(1-\text{SSIM}+\text{LPIPS})} \tag{15}$$

where $(1-\text{SSIM})$ and LPIPS serve as two fidelity-related error terms: the former penalizes structural dissimilarity and the latter penalizes perceptual distortion. Equal weights are used as a simple default because no task-specific preference between structural and perceptual fidelity is assumed in this study. The exponential form monotonically maps larger fidelity errors to smaller positive scores and avoids hard clipping when LPIPS is not strictly bounded by one. Thus, $FI \in (0,1]$, reaching 1 only when SSIM $= 1$ and LPIPS $= 0$. FCSI is then defined as the harmonic mean of CSI and FI:

$$\text{FCSI} = 2 \times \frac{\text{CSI} \times FI}{\text{CSI} + FI} \tag{16}$$

The harmonic mean is chosen because it is more sensitive to the smaller component than an arithmetic mean. FCSI is computed separately at each reflectivity threshold by using the corresponding CSI value at 10, 20, or 30 dBZ. Operationally, a higher FCSI indicates that the model maintains both reasonable threshold-based overlap and visual fidelity under the same threshold setting. A high value in only one component cannot fully compensate for a low value in the other. This property follows the same balancing rationale as F-score-type statistics and makes FCSI useful for analyzing the empirical accuracy-fidelity trade-off observed in this study.

4.3. Comparison Experiment

To comprehensively evaluate the overall performance of models, the proposed RadarEchoMamba is designed in two variants with different parameter scales (RadarEchoMamba-Light, RadarEchoMamba-Base) and compared with representative models from three paradigms: RNNs, Transformers, and CNNs. Specifically, among RNN-based models, we selected the pioneering ConvLSTM and PredRNN, which are representative recurrent base-

lines. Notably, we also included VMRNN, which employs Mamba modules but still adopts the traditional recursive inference mode, to verify the advantages of our end-to-end architecture. Among Transformer-based models, we compared Rainformer and Earthformer, which are designed specifically for spatiotemporal prediction, as well as MS-RadarFormer, which performs strongly on pixel-level metrics. Furthermore, we selected SimVP, known for its high efficiency, as the representative of CNN-based architectures. These baseline models cover the major radar extrapolation paradigms and support the breadth of the comparative experiments. The metrics of different models on the Shanghai-2020 and South China datasets are shown in Tables 1 and 2. These metrics capture different aspects of performance, CSI and POD quantify threshold-based overlap and event detection after binarizing reflectivity fields, whereas SSIM and LPIPS evaluate structural similarity and perceptual distortion in continuous radar images. Table 1 shows this distinction on the Shanghai-2020 dataset. MS-RadarFormer obtains higher CSI values at all three thresholds, whereas RadarEchoMamba-Light obtains better FAR, SSIM, and LPIPS values. Table 2 shows a similar metric-dependent pattern on the South China dataset. MS-RadarFormer achieves the highest CSI and POD values across the three thresholds, indicating strong threshold-based detection. In contrast, RadarEchoMamba-Base achieves the best SSIM and LPIPS, indicating stronger structural and perceptual fidelity. Therefore, the main advantage of RadarEchoMamba in these comparisons lies in visual-fidelity metrics, while its threshold-based accuracy is competitive.

Table 1. Quantitative comparison of different models under different thresholds on Shanghai-2020 Dataset. The best results are highlighted in bold. ↑ (↓) denotes higher (lower) is better.

Algorithm	CSI ↑			FAR ↓			SSIM ↑	LPIPS ↓
	10 dBZ	20 dBZ	30 dBZ	10 dBZ	20 dBZ	30 dBZ		
RadarEchoMamba-Light	0.5316	0.4774	0.3320	**0.2399**	**0.2297**	**0.4063**	**0.8619**	**0.1905**
MS-RadarFormer	**0.5922**	**0.5605**	**0.3969**	0.2663	0.2779	0.4902	0.8138	0.2412
ConvLSTM	0.5664	0.5093	0.3169	0.2178	0.2222	0.3632	0.8395	0.2436

Table 2. Quantitative comparison of different models under different thresholds on South China dataset. The best results are highlighted in bold and the second-best results are underlined. ↑ (↓) denotes higher (lower) is better.

Algorithm	CSI ↑			POD ↑			FAR ↓			SSIM ↑	LPIPS ↓
	10 dBZ	20 dBZ	30 dBZ	10 dBZ	20 dBZ	30 dBZ	10 dBZ	20 dBZ	30 dBZ		
RadarEchoMamba-Light	0.6497	0.6122	0.5186	0.7809	0.6944	0.6348	0.2054	<u>0.1620</u>	0.2609	0.8738	0.1875
RadarEchoMamba-Base	0.6024	0.5791	0.4836	0.6890	0.6620	0.6057	**0.1726**	0.1778	0.2943	**0.8946**	**0.1611**
MS-RadarFormer	**0.6525**	**0.6435**	**0.5490**	**0.8217**	**0.7826**	**0.7424**	0.2399	0.2165	0.3218	0.8142	0.2626
Rainformer	<u>0.6514</u>	<u>0.6257</u>	0.4925	<u>0.7858</u>	<u>0.7293</u>	0.6257	0.2080	0.1851	0.3018	<u>0.8819</u>	<u>0.1818</u>
Earthformer	0.6492	0.6189	<u>0.5309</u>	0.7788	0.7017	<u>0.6442</u>	<u>0.2040</u>	**0.1600**	<u>0.2488</u>	0.8666	0.1911
ConvLSTM	0.6433	0.6147	0.5185	0.7771	0.7085	0.6344	0.2112	0.1772	0.2606	0.8731	0.2184
PredRNN	0.6336	0.6018	0.5015	0.7642	0.6943	0.6060	0.2124	0.1813	0.2559	0.8783	0.1928
VMRNN	0.6393	0.6102	0.5176	0.7654	0.6958	0.6159	0.2050	0.1677	**0.2357**	0.8728	0.1906
SimVP	0.5857	0.5435	0.4106	0.7414	0.6343	0.5018	0.2640	0.2084	0.3068	0.8500	0.2376

Table 3 summarizes the computational profiles of all evaluated models under the same input setting. Combined with Tables 1 and 2, RadarEchoMamba-Light is the efficiency-oriented variant of the proposed framework. It uses 32.327 M parameters, corresponding to 20.8% of the parameter count of MS-RadarFormer, and achieves a lower measured inference latency (0.0715 s vs. 0.0950 s). Although the FLOPs of RadarEchoMamba-Light are slightly higher than those of MS-RadarFormer (397.52 G vs. 343.29 G), the model has a

much smaller parameter count and lower measured inference latency. Table 3 also shows that computational cost is not directly proportional to prediction fidelity. SimVP is the most economical baseline in terms of FLOPs and latency, but its SSIM and LPIPS results in Table 2 are weaker than those of RadarEchoMamba. VMRNN has fewer parameters and FLOPs than RadarEchoMamba-Light, yet its recurrent inference leads to higher measured latency and slightly weaker SSIM/LPIPS performance. ConvLSTM has a parameter scale comparable to RadarEchoMamba-Light, but requires substantially higher FLOPs and latency. These comparisons indicate that RadarEchoMamba-Light occupies a practical operating point with a compact parameter scale, low measured latency, and favorable visual-fidelity metrics. RadarEchoMamba-Base, in contrast, is a capacity-oriented variant used to examine the fidelity attainable with increased model scale, rather than to support the real-time efficiency claim.

The following visual cases complement the tabular results by showing how metric differences appear in the predicted echo fields. We focus on advection trajectory, convective-core compactness, weak-boundary texture, and spatial dilation, because these attributes correspond to the threshold-based and visual-fidelity metrics discussed above.

Table 3. Comparison of computational efficiency among different models. We report the number of parameters, floating point operations (FLOPs), and inference latency for one inference step.↓ denotes lower is better.

Algorithm	Parameters ↓	FLOPs ↓	Latency ↓
RadarEchoMamba-Light	32.327 M	397.52 G	0.0715 s
RadarEchoMamba-Base	241.157 M	1711.90 G	0.1925 s
MS-RadarFormer	155.617 M	343.29 G	0.0950 s
ConvLSTM	27.581 M	722.20 G	0.6258 s
VMRNN	9.985 M	292.80 G	0.1788 s
SimVP	14.421 M	5.10 G	0.0060 s

All models are computed under same input dimensions: $B = 1, T = 20, C = 1, H = W = 512$.

Figure 5 visualizes the evolution of a large-scale convective system characterized by high-intensity cores, providing intuitive validation of the generalization capability of the RadarEchoMamba architecture. RadarEchoMamba tracks the macroscopic advection trend and retains irregular fine-grained textures and compact morphological structures of convective cores in this case. MS-RadarFormer produces broader high-intensity regions that overlap strongly with the observed convective core. Although this spatial dilation strategy artificially boosts CSI metrics by maximizing the prediction coverage, the resulting imagery deviates from the ground truth in structure and physical realism.

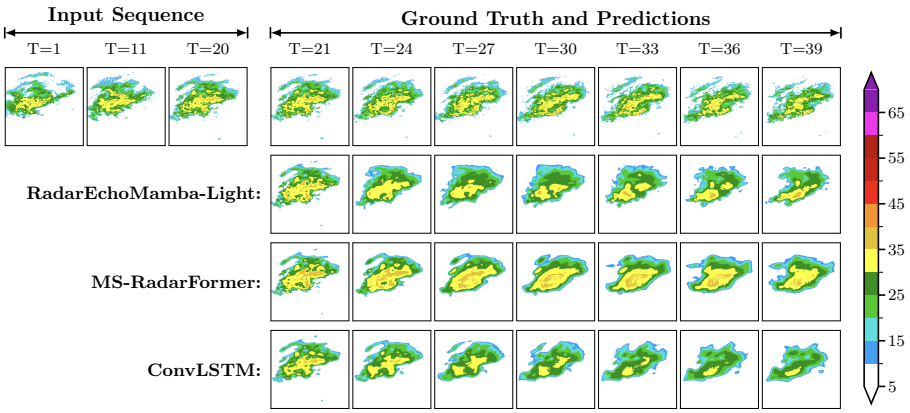

Figure 5. Visualization comparison of an example of Shanghai-2020 dataset. This case illustrates the evolution of a large-scale system with high-intensity cores. The first row shows selected input frames at T = 1, 11, and 20, while subsequent rows display selected predicted frames at T = 21, 24, 27, 33, 36, and 39, alongside the corresponding ground truth. All frames share a uniform color scale in dBZ.

Figure 6 illustrates the extrapolation of a convective system moving towards the northeast. RadarEchoMamba-Light uses 32.327 M parameters, corresponding to 20.8% of the parameter count of MS-RadarFormer (155.617 M), while maintaining competitive structural preservation. It accurately predicts the advection trajectory and preserves the fine-grained texture at the lower right of the main echo body even after extrapolation for two hours (T = 39). This visual evidence is consistent with its LPIPS (0.1875) score in Table 2, indicating favorable high-frequency detail reconstruction by the non-autoregressive Mamba architecture. Transformer-based models like MS-RadarFormer and Earthformer, while performing well in maintaining the intensity of high-echo regions, suffer from significant over-smoothing. Their predictions exhibit distinct blocky artifacts with blurred edges, failing to reconstruct natural radar textures, which leads to poorer perceptual quality. Recursive models (ConvLSTM, PredRNN) exhibit substantial intensity attenuation in the later stages. VMRNN, which employs Mamba modules but relies on a recursive strategy, shows a similar attenuation pattern. As prediction time increases, the echoes progressively dissipate, making it difficult to maintain the morphology of the convective system. SimVP, constrained by its lack of long-range temporal modeling, shows less clear structural reconstruction from the early stages. In summary, RadarEchoMamba-Light achieves a favorable balance between meteorological consistency and visual fidelity under a compact parameter budget.

Figure 7 displays a more complex convective system involving the movement of high-echo regions and non-linear shape deformation. Throughout the two-hour prediction, the RadarEchoMamba series maintained clear structural details. RadarEchoMamba-Base, with its larger model capacity, generated clearer texture details than the other evaluated methods in this example. This observation aligns with its higher SSIM (0.8946) and lower LPIPS (0.1611) scores in Table 2, indicating effective capture of micro-scale convective structures. RadarEchoMamba-Light also captured the morphological deformation and the dissipation trend of the weak echo edge at the bottom right. The Light model retained competitive visual fidelity while using 32.327 M parameters, which is 86.6% fewer than RadarEchoMamba-Base (241.157 M). In comparison, although MS-RadarFormer captured the macroscopic motion through attention mechanisms, it tended to over-smooth the echo fields. Specifically, it predicted the echo center as large, blurred blocks of high intensity. This strategy can improve CSI and POD by expanding the prediction area, but it may also increase FAR and reduce high-frequency texture fidelity. SimVP and VMRNN tended to generate connected and blurry prediction fields, making high-intensity cores less distin-

guishable. Meanwhile, RNN-based models (ConvLSTM, PredRNN) showed substantial intensity attenuation in edge regions and increasingly blurred morphological features due to long-term error accumulation. This comparison indicates that RadarEchoMamba-Light retained clearer multi-scale structural details in this complex example under a compact parameter budget.

To provide an auxiliary summary of the balance between threshold-based accuracy and visual fidelity, we present the FI and FCSI results in Table 4. FCSI is used here as a supplementary statistic to be read together with the original metrics in Table 2. It summarizes whether a model maintains overlap accuracy without a large degradation in structural or perceptual fidelity. Although MS-RadarFormer exhibits strong CSI and POD scores in Table 2, its lower SSIM and higher LPIPS lead to a lower Fidelity Index ($FI = 0.6386$). Consequently, it receives lower FCSI scores across all thresholds (e.g., 0.5904 at 30 dBZ). Similarly, traditional recursive models such as ConvLSTM and SimVP show a poorer balance between accuracy and fidelity.

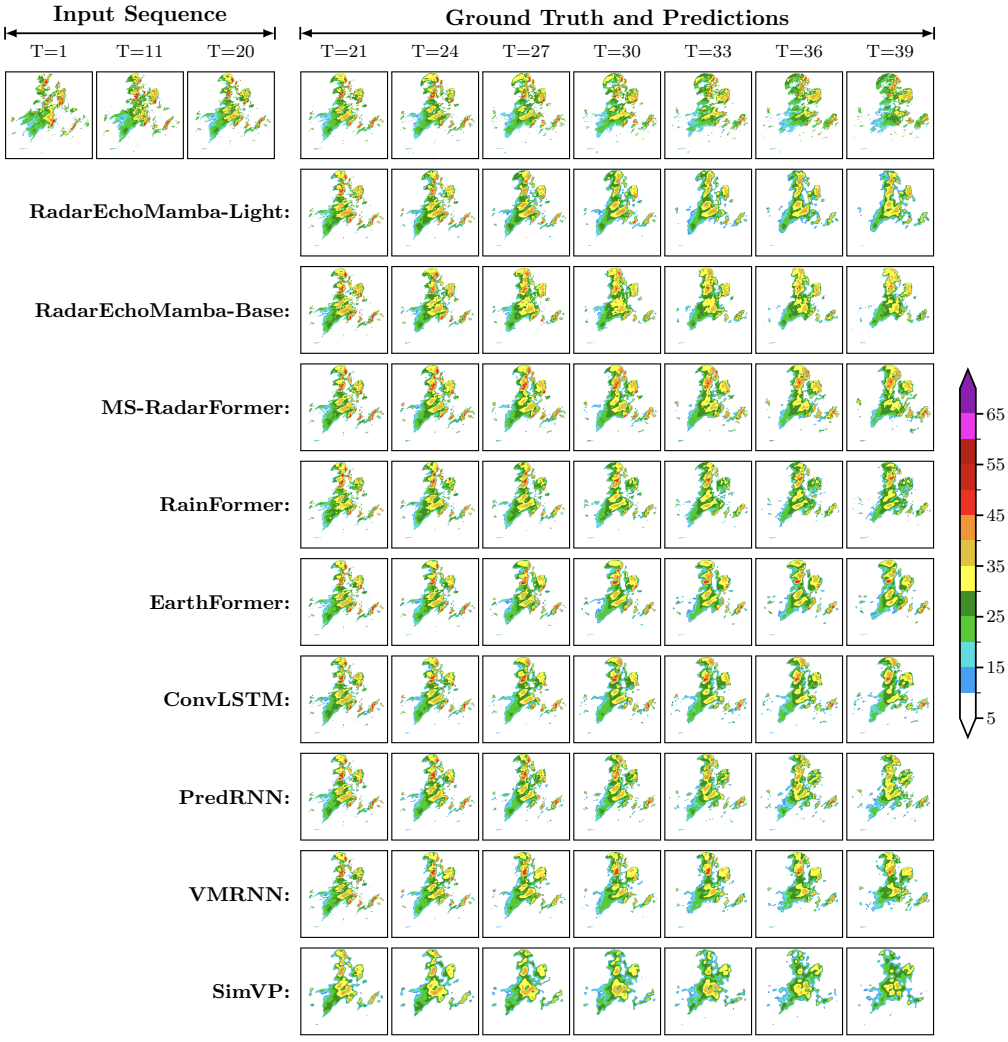

Figure 6. Visualization comparison of a convective system moving northeast on the South China dataset. This case shows the non-linear deformation and intensity evolution of a high-intensity core over two hours. The first row shows selected input frames at T = 1, 11, and 20, while subsequent rows display selected predicted frames at T = 21, 24, 27, 33, 36, and 39, alongside the corresponding ground truth. All frames share a uniform color scale in dBZ.

Table 4. FCSI comparison on the South China dataset. ↑ denotes higher is better. The best and second-best results are highlighted in bold and underlined.

Algorithm	*FI* ↑	FCSI ↑		
		10 dBZ	**20 dBZ**	**30 dBZ**
RadarEchoMamba-Light	<u>0.7307</u>	**0.6878**	**0.6662**	**0.6067**
RadarEchoMamba-Base	**0.7661**	0.6744	0.6596	0.5929
MS-RadarFormer	0.6386	0.6455	0.6411	0.5904
ConvLSTM	0.7080	0.6741	0.6581	0.5986
VMRNN	0.7277	<u>0.6807</u>	<u>0.6638</u>	<u>0.6049</u>
SimVP	0.6787	0.6288	0.6036	0.5117

The RadarEchoMamba series obtains favorable FI and FCSI values. The Base variant achieves the highest Fidelity Index ($FI = 0.7661$), which is 0.1275 higher than MS-RadarFormer and 0.0384 higher than VMRNN. Meanwhile, the Light variant obtains the highest FCSI values across the three thresholds, reaching 0.6878, 0.6662, and 0.6067 at 10, 20, and 30 dBZ, respectively. The absolute margins over the second-best FCSI values are 0.0071, 0.0024, and 0.0018, respectively; therefore, these results are interpreted as a small but consistent advantage in the proposed auxiliary composite metric rather than as a large performance gap.

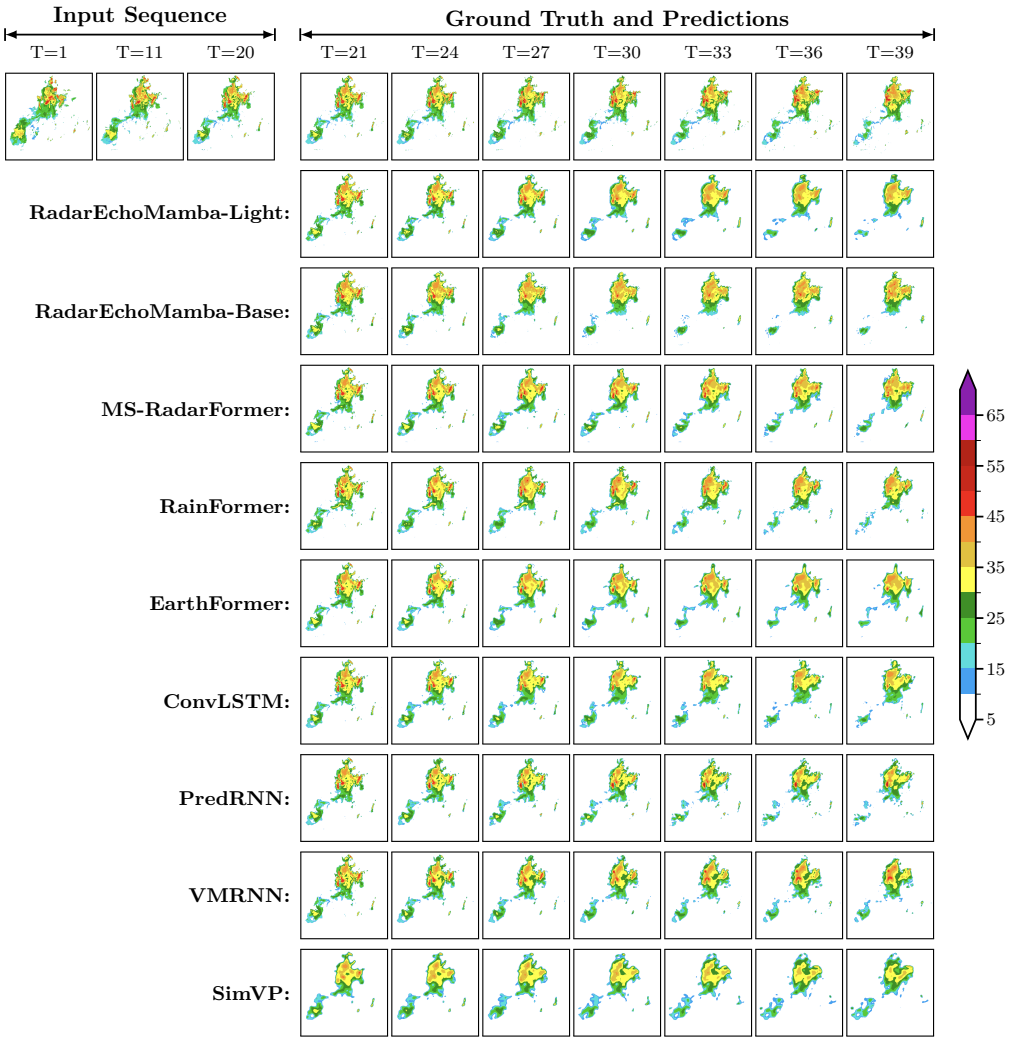

Figure 7. Visualization comparison of a complex evolutionary process on the South China dataset. This case shows morphological deformation and the dissipation of weak boundary echoes over two hours. The first row shows selected input frames at T = 1, 11, and 20, while subsequent rows display selected predicted frames at T = 21, 24, 27, 33, 36, and 39, alongside the corresponding ground truth. All frames share a uniform color scale in dBZ.

These results suggest that our pure Mamba-based architecture mitigates the inherent accuracy-fidelity dilemma, offering a balance between spatial precision and visual realism. By leveraging the linear-complexity bidirectional SSM and the context-guided decoder, RadarEchoMamba reduces the geometric dilation typically induced by MSE loss, providing high-precision advection tracking while preserving the physical structural integrity of meteorological systems. We further compared the paired time-step differences in SSIM and LPIPS in Figure 8j,k. RadarEchoMamba-Base obtains higher SSIM than MS-RadarFormer and Earthformer at all 20 lead times, and higher SSIM than Rainformer at 16 of the 20 lead times. For LPIPS, it achieves lower values than all three Transformer baselines at every lead time. A one-sided paired sign test further indicates that these trends are consistent across the prediction horizon ($p \leq 0.0059$). Therefore, the SSIM/LPIPS gains are interpreted as a stable visual-fidelity trend rather than as improvements caused by a few isolated lead times. Taken together, these results show that RadarEchoMamba provides more consistent structural and perceptual fidelity, especially in preserving sharper echo boundaries and more coherent convective morphology over longer prediction horizons.

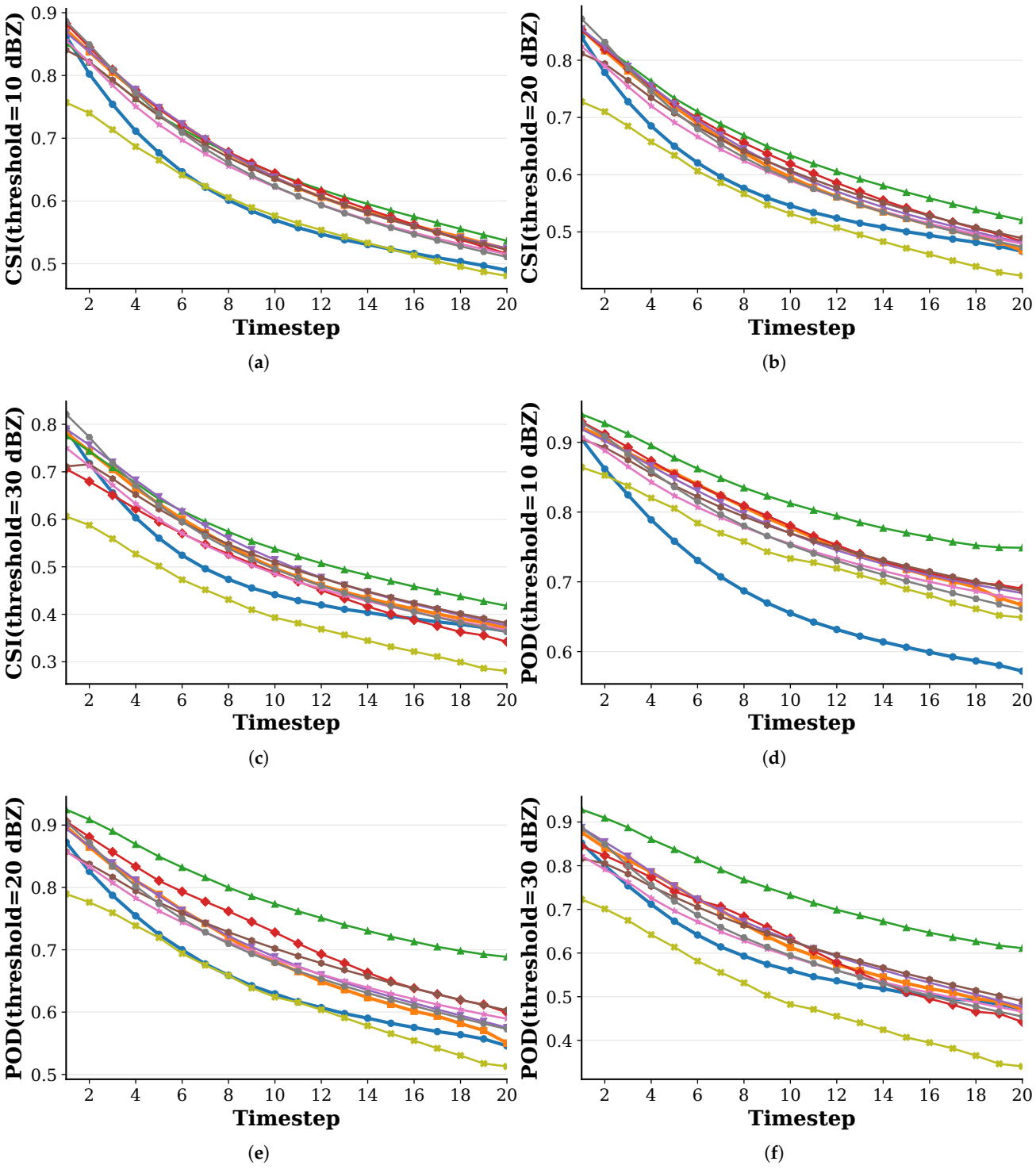

Figure 8. *Cont.*

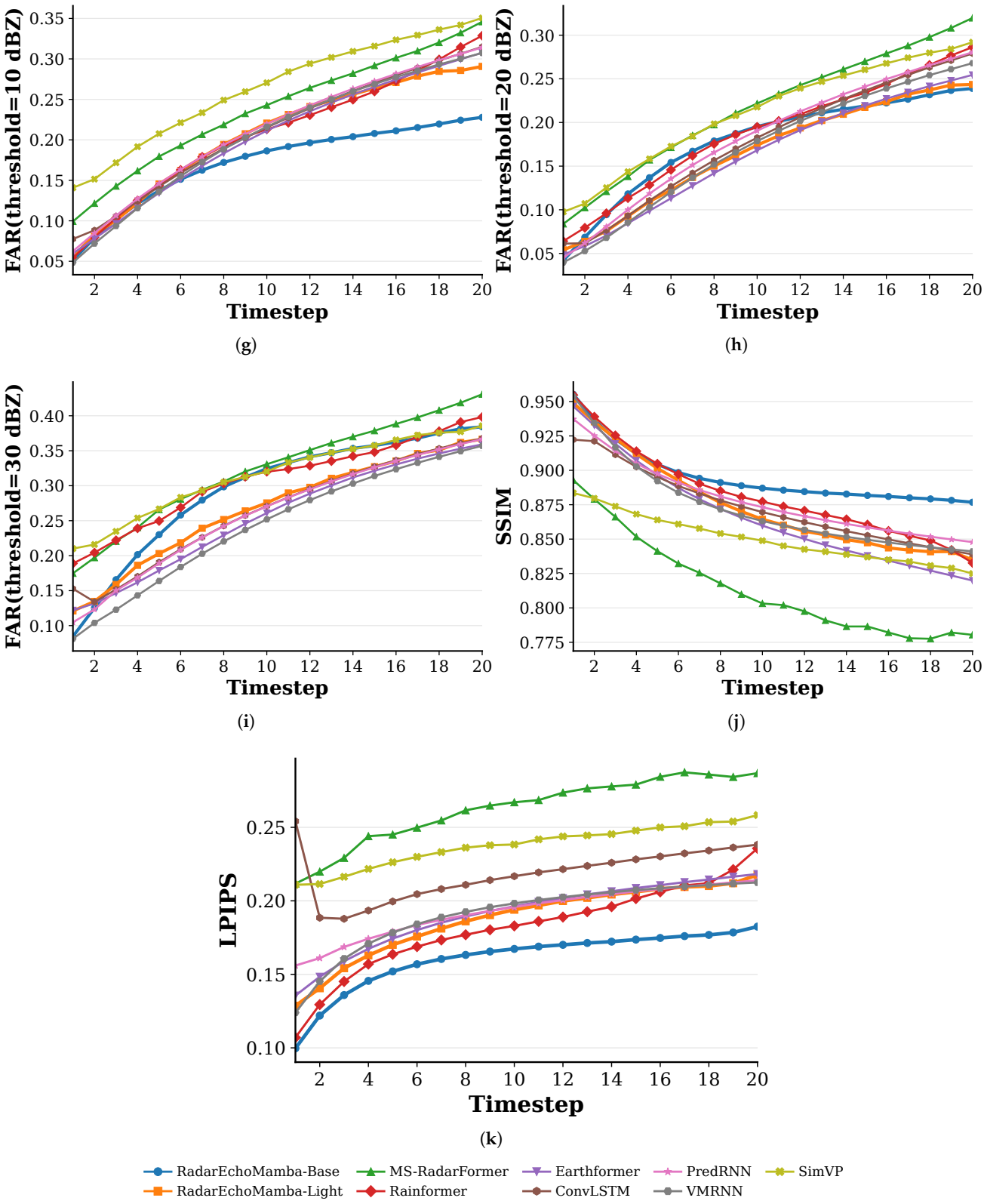

Figure 8. Performance evolution curves of different models over 20 prediction time steps on the South China dataset. (**a**) CSI at 10 dBZ. (**b**) CSI at 20 dBZ. (**c**) CSI at 30 dBZ. (**d**) POD at 10 dBZ. (**e**) POD at 20 dBZ. (**f**) POD at 30 dBZ. (**g**) FAR at 10 dBZ. (**h**) FAR at 20 dBZ. (**i**) FAR at 30 dBZ. (**j**) Structural Similarity Index (SSIM). (**k**) Learned Perceptual Image Patch Similarity (LPIPS). The x-axis denotes the prediction time steps.

4.4. Ablation Study

To systematically evaluate the effectiveness and contribution of each core component within the RadarEchoMamba framework, we conducted a series of ablation studies. To ensure fairness and consistency, all ablation variants were built on a baseline model configured with an embedding dimension of 768 and a depth of 8 layers.

4.4.1. Core Components

As shown in Table 5, we analyze the impact of key modules by progressively removing them. The full model achieves the best 30 dBZ CSI and POD and the lowest LPIPS, whereas the differences at lower thresholds are comparatively small. This pattern indicates that the modules mainly affect severe-echo evolution and structural reconstruction, rather than producing uniform gains across all metrics. Therefore, the high-threshold scores are used to examine convective-core preservation, while SSIM, LPIPS, and the visual examples are used to assess boundary sharpness and perceptual fidelity.

Table 5. Quantitative comparison of modules. CSI, POD, and FAR are computed at 10, 20, and 30 dBZ thresholds. SSIM and LPIPS evaluate structural and perceptual similarity, respectively. ↑ (↓) denotes higher (lower) is better. The best and second-best results are highlighted in bold and underlined.

Algorithm	CSI ↑			POD ↑			FAR ↓			SSIM ↑	LPIPS ↓
	10 dBZ	20 dBZ	30 dBZ	10 dBZ	20 dBZ	30 dBZ	10 dBZ	20 dBZ	30 dBZ		
w/o MSTE	**0.6040**	**0.5796**	0.4783	<u>0.6933</u>	**0.6645**	<u>0.5951</u>	0.1758	0.1807	0.2909	0.8944	<u>0.1624</u>
w/o bidirectional	<u>0.6029</u>	0.5763	<u>0.4788</u>	**0.6977**	0.6614	0.5933	0.1839	0.1824	<u>0.2873</u>	0.8906	0.1681
w/o conditional decoder	0.5999	0.5726	0.4767	0.6818	0.6480	0.5850	**0.1669**	**0.1688**	**0.2796**	<u>0.8946</u>	0.1629
RadarEchoMamba-Base	0.6024	<u>0.5791</u>	**0.4836**	0.6890	<u>0.6620</u>	**0.6057**	<u>0.1726</u>	<u>0.1778</u>	0.2943	**0.8946**	**0.1611**

Removing the bidirectional mechanism degrades both perceptual quality and high-threshold detection. LPIPS increases from 0.1611 to 0.1681, and the 30 dBZ CSI/POD decrease from 0.4836/0.6057 to 0.4788/0.5933. Since the 30 dBZ threshold mainly reflects intense convective echoes, this degradation indicates that forward and time-reversed temporal contexts help stabilize the evolution of strong echo regions while maintaining visual fidelity.

The removal of the MSTE module produces mixed changes at low thresholds but weakens high-threshold detection. On the South China dataset, the 30 dBZ CSI and POD decrease from 0.4836 and 0.6057 to 0.4783 and 0.5951, respectively. This result indicates that multi-scale tubelet embeddings support the representation of convective cores and scale-dependent echo morphology. Because MSTE contains fine, medium, and macroscopic branches, its contribution is related not only to intense echo regions but also to the surrounding weak-boundary context.

Removing the conditional decoder lowers CSI and POD at all thresholds while reducing FAR. This indicates that the predictions become more conservative, with fewer false alarms but more missed echo areas. LPIPS also worsens from 0.1611 to 0.1629. Since the decoder introduces input-derived structural priors during progressive reconstruction, this result is consistent with its role in weak boundary recovery and visual fidelity. Overall, the full configuration achieves the best 30 dBZ CSI and POD and the lowest LPIPS among the tested variants, while matching the best SSIM. The limited absolute gaps indicate that these modules act jointly and that their effects differ across threshold-based and visual-fidelity metrics. These module-level ablations are further interpreted together with the diagnostic comparison among Forward Only, Backward Only, Arithmetic Average, and the proposed learnable fusion in Section 5.

4.4.2. Cross-Dataset Stability Verification

To further verify the robustness and reproducibility of the above ablation conclusions, we re-evaluate all ablation variants on the independent Shanghai-2020 dataset with three independent training runs and report the mean and standard deviation in Table 6. As shown, the standard deviations across all variants remain consistently low (typically below 0.01), confirming the training stability of the proposed framework. In terms of metric performance, the Light version of the full model (RadarEchoMamba-Light) achieves the best CSI (0.3300 ± 0.0022) and POD (0.4244 ± 0.0054) at the most critical severe convective threshold of 30 dBZ, achieving the highest values among the ablation variants. Although variants without MSTE or the conditional decoder yield marginally higher SSIM and lower LPIPS scores, they consistently suffer from notable degradation in high-threshold detection accuracy. This observation aligns with the accuracy-fidelity trade-off discussed in the following section: upon removing structural priors or multi-scale embedding, the model tends to produce smoother predictions that achieve better perceptual scores at the expense of precise detection of high-intensity echo regions. Overall, the full model achieves a favorable balance between threshold-based accuracy and perceptual fidelity across the two evaluated datasets.

Table 6. Cross-dataset stability verification of the ablation study on the Shanghai-2020 dataset. Results are reported as mean ± std over three independent runs. ↑ (↓) denotes higher (lower) is better. The best and second-best results are highlighted in bold and underlined.

Algorithm	CSI ↑			POD ↑			SSIM ↑	LPIPS ↓
	10 dBZ	20 dBZ	30 dBZ	10 dBZ	20 dBZ	30 dBZ		
w/o MSTE	0.5164 ± 0.0047	**0.4771 ± 0.0020**	0.3145 ± 0.0073	0.5973 ± 0.0100	0.5557 ± 0.0057	0.3969 ± 0.0162	**0.8804 ± 0.0023**	**0.1782 ± 0.0016**
w/o bidirectional	**0.5230 ± 0.0060**	0.4720 ± 0.0086	0.2993 ± 0.0082	**0.6393 ± 0.0190**	**0.5657 ± 0.0262**	0.3714 ± 0.0197	0.8628 ± 0.0054	0.1906 ± 0.0026
w/o conditional decoder	0.5078 ± 0.0123	0.4644 ± 0.0032	0.3137 ± 0.0156	0.5906 ± 0.0241	0.5400 ± 0.0068	0.4069 ± 0.0511	0.8760 ± 0.0041	0.1822 ± 0.0038
RadarEchoMamba-Light	0.5197 ± 0.0142	0.4721 ± 0.0071	**0.3300 ± 0.0022**	0.6092 ± 0.0355	0.5460 ± 0.0164	**0.4244 ± 0.0054**	0.8714 ± 0.0093	0.1839 ± 0.0064

4.4.3. Model Scaling

We further explore the impact of model depth and width (embedding dimension), as shown in Table 7. While maintaining an embedding dimension of 768, scaling the depth from 8 to 16 layers reveals a clear trend of continuous improvement in SSIM and LPIPS. Although pixel-level overlap metrics (CSI) are affected by the double-penalty effect associated with sharper edges, the physical fidelity of the predictions improves. RadarEchoMamba-Light achieves the highest CSI and POD scores but yields the poorest perceptual quality among the RadarEchoMamba variants. This phenomenon is consistent with the observation that models optimized for higher CSI may tend to produce smoother predictions.

Table 7. Impact of embedding dimension and network depth on performance. The best results are highlighted in bold and the second-best results are underlined. ↑ (↓) denotes higher (lower) is better.

Embed Dim	Depth	CSI ↑			POD ↑			FAR ↓			SSIM ↑	LPIPS ↓
		10 dBZ	20 dBZ	30 dBZ	10 dBZ	20 dBZ	30 dBZ	10 dBZ	20 dBZ	30 dBZ		
768	8	0.6024	0.5791	0.4836	0.6890	0.6620	0.6057	0.1726	0.1778	0.2943	0.8946	0.1611
768	12	0.5908	0.5677	0.4771	0.6689	0.6444	0.5930	0.1651	0.1734	0.2905	0.8948	0.1605
768	16	0.5917	0.5661	0.4717	0.6702	0.6424	0.5857	**0.1651**	0.1733	0.2921	**0.8955**	**0.1595**
256	8	**0.6497**	**0.6122**	**0.5186**	**0.7809**	**0.6944**	**0.6348**	0.2054	**0.1620**	**0.2609**	0.8738	0.1875

5. Discussion

5.1. Trade-Off Between Threshold Metrics and Image Metrics

Based on the scatter plot analysis in Figure 9, we observe a negative association between threshold-based pixel-level accuracy and perceptual or structural fidelity among

the evaluated models. The gray dashed line connecting the major models outlines this empirical trend: as a model moves toward the bottom right region with higher CSI, its perceptual quality often decreases along the vertical axis.

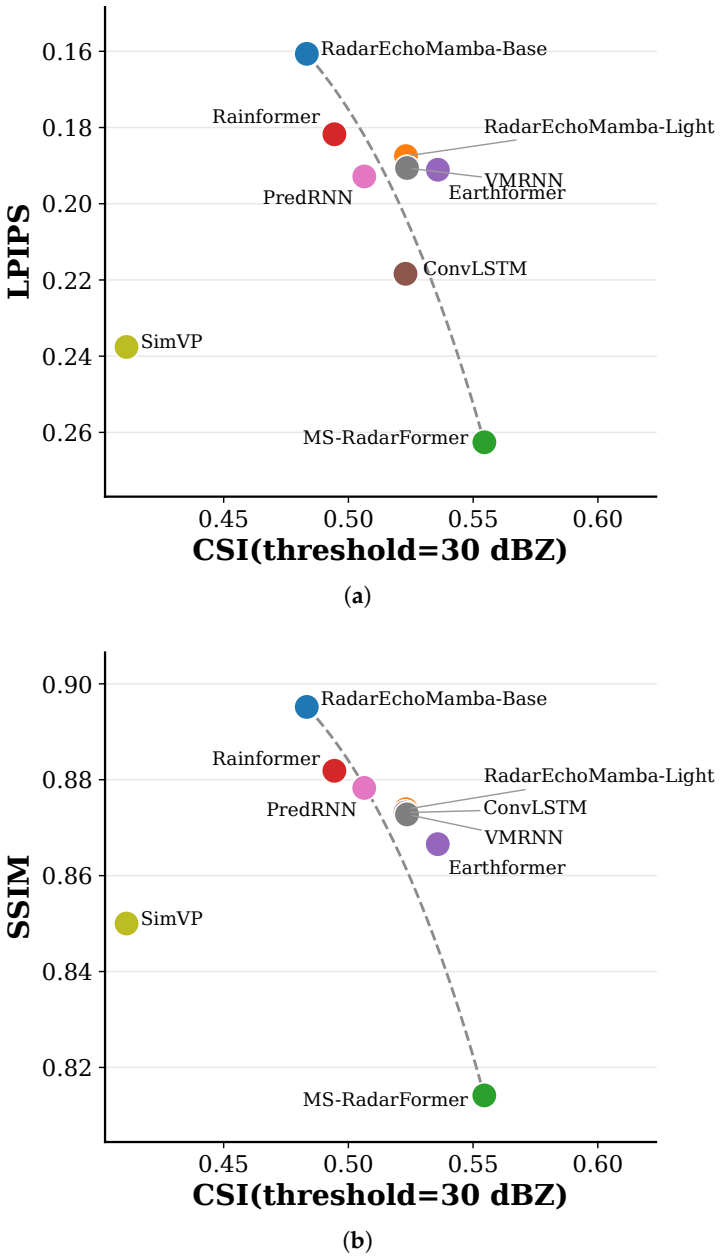

Figure 9. Scatter plots illustrating the relationship between threshold-based accuracy (CSI) and perceptual/structural fidelity metrics. (**a**) CSI vs. LPIPS. (**b**) CSI vs. SSIM. The gray dashed line indicates the inverse correlation trend among models.

Models tend to generate blurry prediction results with expanded spatial extent. Although this behavior can increase CSI by enlarging the pixel overlap area with the ground truth, it may lose high-frequency texture information, leading to poorer performance in LPIPS and SSIM scores. By comparison, RadarEchoMamba lies near the top-left end of the dashed line, showing better perceptual quality through predictions with sharper edges and richer details.

However, in the presence of spatiotemporal uncertainty, sharpening high-frequency details increases the risk of minor spatial misalignments. This can trigger the double-penalty effect of threshold-based metrics and result in lower CSI scores. This empirical

observation shows why a single metric is insufficient for describing both detection accuracy and structural fidelity in deterministic nowcasting.

This trade-off phenomenon is not isolated. The Earthformer paper also pointed out that model variants with better skill scores have lower perceptual similarity between generated results and real images. Conversely, visually more satisfactory predictions may be accompanied by poorer skill scores. This finding is consistent with our observation that optimizing pixel-level metrics in deterministic models may drive predictions toward smoothing and blurring, thereby weakening physical structural fidelity.

5.2. Qualitative Analysis of the Trade-Off

To investigate the possible explanation of the accuracy-fidelity trade-off, this paper attempts to conduct a qualitative analysis based on the optimization objective of the loss function and the definition of SSIM. Mainstream loss function, during the optimization process, may encourage the model to approximate the conditional expectation of the future state $\hat{Y} = \mathbb{E}_{Y|X}[Y]$ as Equation (19). Under the inherent uncertainty of weather systems, the future state Y often does not follow a specific distribution, and the process of taking the expectation of the future state is equivalent to resulting in an averaging of the model's prediction, which leads to predict a blurred, expanded echo blob covering all probable trajectories. While this strategy can artificially maintain high global variance and boost CSI, it comes at the cost of local structural integrity. This dilation destroys geometric topology: real echoes are fractal and sharp, whereas predictions become rounded and diffuse. This mismatch significantly reduces the local covariance between prediction and ground truth. According to Equation (17), SSIM is governed by covariance between Y and $\hat{Y}$. Consequently, the smoothing and dilation induced by MSE inevitably lead to a drop in $\sigma_{\hat{Y}Y}$, degrading the SSIM score.

To examine the above theory, we contrast the Global Variance $\sigma(\hat{Y})$ with the structure component s of SSIM across MS-RadarFormer and RadarEchoMamba-Light. The data shows that MS-RadarFormer maintains a high global variance ($\sigma = 0.0132$) compared with RadarEchoMamba ($\sigma = 0.0111$). In contrast, while RadarEchoMamba exhibits a slight convergence in global variance, it achieves a superior structure score of 0.945, MS-RadarFormer get 0.938 in structure score, representing an 11.3% reduction in relative structural error.

Taken together, the higher global variance and lower structure score of MS-RadarFormer suggest that part of its variance may come from spatially expanded and diffuse echo regions rather than from accurately localized structures. Such predictions can increase CSI by enlarging the overlap with the observed echo area, but they may also weaken local geometric consistency and blur high-frequency boundaries. In contrast, RadarEchoMamba achieves a higher structure score despite a slightly lower global variance, indicating that its predicted variations are more consistent with the observed echo morphology. This comparison supports the interpretation that visual fidelity depends not only on the magnitude of predicted variability, but also on whether this variability corresponds to physically plausible echo structures.

$$s(\hat{Y}, Y) = \frac{\sigma_{\hat{Y}Y} + C_3}{\sigma_{\hat{Y}}\sigma_Y + C_3} \tag{17}$$

$$\begin{aligned}\mathcal{L} &= \frac{1}{N}\sum_{i=1}^{N}(Y-\hat{Y})^2 \\ &= \mathbb{E}_{Y|X}\left[(Y-\hat{Y})^2\right] \\ \frac{d\mathcal{L}}{d\hat{Y}} &= \frac{d}{d\hat{Y}}\mathbb{E}_{Y|X}\left[(Y-\hat{Y})^2\right] \\ &= \mathbb{E}_{Y|X}\left[\frac{d}{d\hat{Y}}(Y-\hat{Y})^2\right] \\ &= -2\left(\mathbb{E}_{Y|X}[Y]-\hat{Y}\right)\end{aligned} \tag{18}$$

$$\begin{aligned}\mathbb{E}_{Y|X}[Y]-\hat{Y} &= 0 \\ \hat{Y} &= \mathbb{E}_{Y|X}[Y]\end{aligned} \tag{19}$$

5.3. Information Preservation and Feature Evolution in the SequenceEncoder and Backbone

This paper conducted a diagnostic analysis and visual analysis on both the SequenceEncoder and the Bidirectional Mamba Backbone. Our analysis suggests that our architecture suffers less from information decay; rather, it achieves stable information distillation, accumulation, and complementary amplification.

5.3.1. Deliberate Decoupling and Low-Pass Filtering in the SequenceEncoder

A critical concern is whether the temporal pooling operation in the SequenceEncoder causes undesirable temporal information loss. We argue that discarding high-frequency temporal dynamics here is a deliberate decoupling strategy. The SequenceEncoder is strictly tasked with providing a static, macroscopic spatial blueprint, while complex temporal dynamics are delegated to the Mamba backbone. Mathematically, the AvgPool operation acts as a temporal low-pass filter. By averaging across the temporal dimension, it suppresses transient dynamic noise and preserves the stable, low-frequency background structure (e.g., the general contour of convective systems). To empirically validate this, we replaced AvgPool with MaxPool in an ablation experiment. The results showed that MaxPool severely degraded the SSIM from 0.8946 to 0.8686 and LPIPS from 0.1611 to 0.1888. This degradation occurs because MaxPool captures the extreme intensity envelope across time, projecting the entire movement trajectory into a single frame. This creates an artificially dilated and smeared spatial prior that misleads the decoder. Therefore, AvgPool is better suited to constructing a stable macroscopic structural prior under the evaluated setting.

5.3.2. Information Accumulation and Refinement in the Mamba Backbone

To investigate whether multi-scale features are lost during deep propagation in backbone, this paper tracks the Mean Feature L2 Norms across different layers, and explicitly decoded intermediate latent features (Layer 1 to Layer 8 for RadarEchoMamba-Base with embed dim = 768 and depth = 8) back into radar echoes.

As the network deepens, the feature norms stably increase without vanishing. Figure 10 shows that shallow layers (Layer 1–Layer 2) retain raw, chaotic spatial multi-scale details but struggle to maintain temporal coherence. Deep layers (Layer 7–Layer 8) successfully filter out high-frequency noise and distill coherent spatiotemporal advection trends, which suggests that information is highly refined.

The feature norms reveal the asymmetric nature of causal modeling. In the Forward path, the norm progressively increases from $t = 0$ to the end of the sequence, as later tokens accumulate more historical context. Conversely, the Backward path peaks at $t = 0$ and decreases forward, accumulating future context. This suggests that a unidirectional

SSM suffers from partial context deficiency at either end of the sequence, necessitating a bidirectional design.

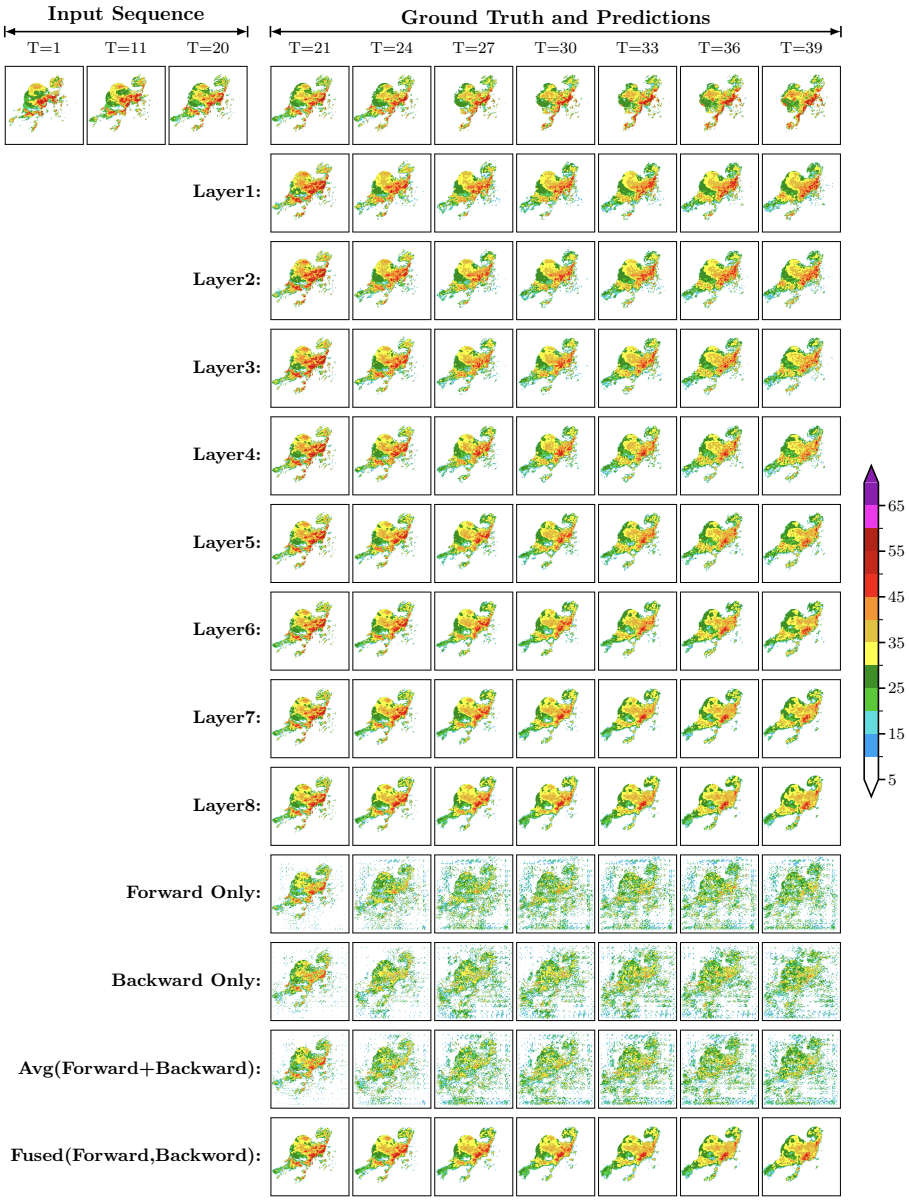

Figure 10. Layer-wise semantic refinement and visual comparison of different temporal fusion strategies based on the RadarEchoMamba-Base model. The top row displays the input sequence and the corresponding ground truth at T = 21, 24, 27, 33, 36, and 39. Layers 1 to 8 illustrate the intermediate latent features explicitly decoded at each corresponding depth, showing the progressive filtering of chaotic spatial noise and the distillation of stable spatiotemporal advection trends. The bottom four rows evaluate the macroscopic impact of different fusion strategies: Forward Only, Backward Only, Arithmetic Average, and our proposed Concat+Linear Fusion. The weights of the first seven layers were completely frozen, and the distinct fusion strategies were applied exclusively at the final layer (Layer 8).

5.3.3. Diagnostic Analysis of Bidirectional Feature Fusion

We further conduct a diagnostic analysis of bidirectional fusion using L2 norms and the visual comparisons in Figures 10 and 11. Let $F_{t,s}$ and $B_{t,s}$ denote the forward and backward feature vectors in the 768-dimensional latent space. As shown in Figure 11,

$\|F_{8,0}\|_2 \approx 57.70$ and $\|B_{8,0}\|_2 \approx 63.45$. However, the norm of their naive average drops to 46.76.

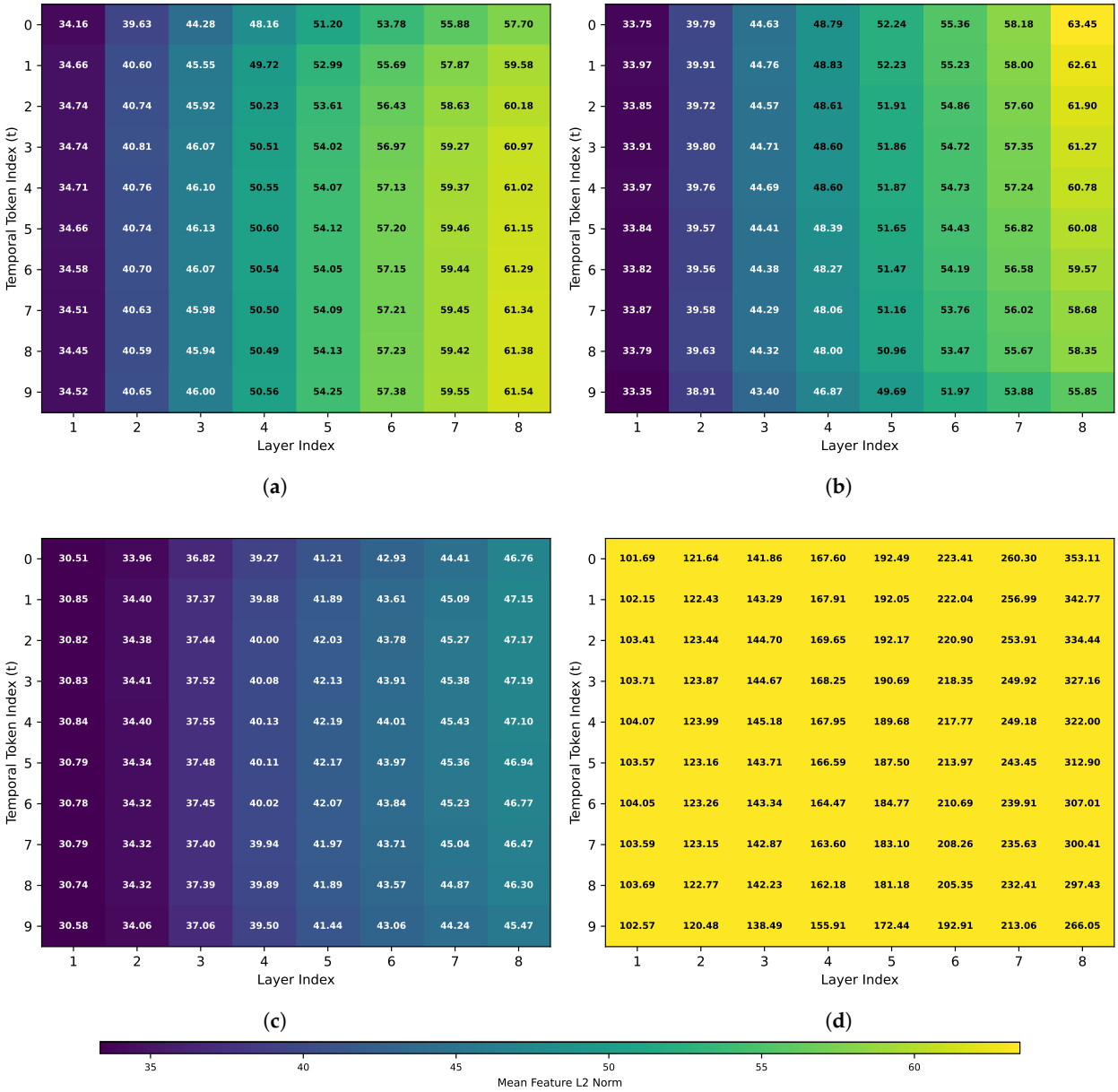

Figure 11. Evolution of Spatial Expected L2 Norms in the Bidirectional Backbone. The heatmaps quantify the average feature activation energy at each layer (X-axis) and temporal token index (Y-axis) for the RadarEchoMamba-Base model. (**a**) Forward path. (**b**) Backward path. (**c**) Arithmetic average. (**d**) Fused representation (Concat + Linear).

According to the triangle inequality in Equation (20), this norm reduction is consistent with partial cancellation between directional features. The norm magnitude alone is insufficient to establish semantic complementarity. However, when combined with the degraded reconstruction produced by simple averaging in Figure 10, it suggests that forward and backward features contain non-identical temporal evidence. In contrast, the learnable projection in the backbone, $W \cdot [F; B]$, avoids the constraint of fixed averaging and produces a fused norm of 353.11. This indicates that the projection actively amplifies

the most informative components from both temporal directions, enabling the decoder to reconstruct remarkably smooth and physically realistic convective structures.

$$\|\frac{F+B}{2}\|_2 \leq \frac{\|F\|_2 + \|B\|_2}{2} \tag{20}$$

5.4. Future Work

Based on these observations and the preceding analysis, this paper suggests that high-resolution radar echo extrapolation can be defined as a spatiotemporal sequence generation problem under multi-objective optimization conditions. The optimization objective can be defined as follows:

$$\min_{\theta} \mathcal{L}_{total} = \lambda_1 \mathcal{L}_{accuracy} + \lambda_2 \mathcal{L}_{fidelity} \tag{21}$$

where $\mathcal{L}_{accuracy}$ denotes the accuracy loss, such as MSE, and $\mathcal{L}_{fidelity}$ represents the fidelity loss, such as SSIM. θ denotes the learnable parameters of the model, and λ_1 and λ_2 are the balancing coefficients.

Most existing methods adopt MSE and its variants as loss functions to optimize the model. According to Equation (21), this practice implicitly assigns a high weight to the accuracy objective and may encourage mean-regression behavior under uncertainty. RadarEchoMamba introduces structural priors through the Multi-Scale Tubelet Embedding and context-guided decoder, which helps preserve fidelity without explicitly adding a perceptual loss. This observation motivates future research. At the algorithmic level, multi-objective optimization algorithms could dynamically balance the weights of accuracy and fidelity losses. Probabilistic diffusion models could also be incorporated to decouple deterministic mean regression from uncertainty representation. At the data level, multi-source meteorological data may reduce the uncertainty of future radar states, improving threshold metrics while preserving structural fidelity.

6. Conclusions

This paper aims to address the problem of prediction blurring and detail distortion in high-resolution radar echo nowcasting. To this end, it proposes RadarEchoMamba, a novel end-to-end extrapolation framework based on the efficient State Space Model (SSM). The core of this framework is a Bidirectional Spatiotemporal Mamba Backbone, which captures long-range spatiotemporal dependencies with linear computational complexity and supports the physical consistency of macroscopic structures. To improve microscopic detail fidelity, this study further designs a Multi-Scale Tubelet Embedding (MSTE) module and a context-guided decoder driven by the original input. The former preserves hierarchical information during encoding, while the latter provides structural priors for detail reconstruction during decoding. Evaluations on the South China and Shanghai-2020 datasets show that RadarEchoMamba achieves favorable structural and perceptual fidelity, especially in SSIM and LPIPS, while maintaining competitive threshold-based metrics. The Light variant further provides a compact and low-latency configuration, whereas the Base variant illustrates the fidelity attainable with increased model capacity. By combining Mamba's efficient global modeling capability with detail-enhancement mechanisms, RadarEchoMamba can generate detail-rich radar echo predictions under the deterministic extrapolation setting evaluated in this study.

Despite the encouraging results achieved by RadarEchoMamba, several avenues remain for future exploration. First, as a deterministic model, it cannot quantify the intrinsic uncertainty associated with the evolution of weather systems. A future direction is to integrate RadarEchoMamba as a conditional backbone with generative models, such as diffusion models, to realize high quality probabilistic forecasting. Second, this study

observes a complex trade-off between meteorological metrics (e.g., CSI) and image quality metrics (e.g., LPIPS). Consequently, exploring novel loss functions or training strategies that can jointly optimize these two categories of metrics remains an important research topic for improving nowcasting performance.

Author Contributions: Conceptualization, Z.S.; methodology, Z.S.; software, F.W.; validation, Z.S., F.W. and H.Z.; formal analysis, Z.S.; investigation, H.G.; resources, J.M.; data curation, H.G.; writing—original draft preparation, Z.S.; writing—review and editing, H.G.; visualization, H.G.; supervision, H.G.; project administration, J.M.; funding acquisition, J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 42375145; the Guangdong S&T Programme, grant number 2024A1111120024; the Major Science and Technology Innovation Demonstration Program of the Inner Mongolia Autonomous Region, grant number 2025ZDSF0009; the Postgraduate Research & Practice Innovation Program of Jiangsu Province, grant number KYCX25_1659; and the Open Grants of China Meteorological Administration Radar Meteorology Key Laboratory, grant number 2025LRM-A03.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT 5.5 for the purposes of English language polishing, improving sentence clarity. The authors have reviewed and edited the output and take full responsibility for the content of this publication.All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SSM	State Space Model
MSTE	Multi-Scale Tubelet Embedding
BSTM	Bidirectional Spatiotemporal Mamba
PD	Progressive Decoder
CSI	Critical Success Index
POD	Probability of Detection
FAR	False Alarm Rate
SSIM	Structural Similarity Index
LPIPS	Learned Perceptual Image Patch Similarity
FCSI	Fidelity-aware Critical Success Index

References

1. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.c. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems 28*; Curran Associates, Inc.: Red Hook, NY, USA, 2015.
2. Wang, Y.; Long, M.; Wang, J.; Gao, Z.; Yu, P.S. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
3. Gao, Z.; Tan, C.; Wu, L.; Li, S.Z. Simvp: Simpler yet better video prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3170–3180.
4. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
5. Tang, Y.; Qi, L.; Xie, F.; Li, X.; Ma, C.; Yang, M.H. Video Prediction Transformers without Recurrence or Convolution. *arXiv* **2024**, arXiv:2410.04733.

6. Gao, Z.; Shi, X.; Wang, H.; Zhu, Y.; Wang, Y.B.; Li, M.; Yeung, D.Y. Earthformer: Exploring space-time transformers for earth system forecasting. In *Advances in Neural Information Processing Systems 35*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; pp. 25390–25403.
7. Geng, H.; Wu, F.; Zhuang, X.; Geng, L.; Xie, B.; Shi, Z. The MS-RadarFormer: A transformer-based multi-scale deep learning model for radar echo extrapolation. *Remote Sens.* **2024**, *16*, 274.
8. Dao, T.; Gu, A. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv* **2024**, arXiv:2405.21060.
9. Shi, X.; Gao, Z.; Lausen, L.; Wang, H.; Yeung, D.Y.; Wong, W.k.; Woo, W.c. Deep learning for precipitation nowcasting: A benchmark and a new model. In *Advances in Neural Information Processing Systems 30*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
10. Wang, Y.; Wu, H.; Zhang, J.; Gao, Z.; Wang, J.; Yu, P.S.; Long, M. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 2208–2225.
11. Geng, H.; Zhao, H.; Shi, Z.; Wu, F.; Geng, L.; Ma, K. MBFE-UNet: A Multi-Branch Feature Extraction UNet with Temporal Cross Attention for Radar Echo Extrapolation. *Remote Sens.* **2024**, *16*, 3956.
12. Li, L.; Zhang, X.; Han, L. SCECA-Net: A Deep Learning-Based Model for Precipitation Nowcasting. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–11.
13. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
14. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv* **2021**, arXiv:2111.00396.
15. Gu, A.; Dao, T.; Ermon, S.; Rudra, A.; Ré, C. Hippo: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems 33*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; pp. 1474–1487.
16. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual state space model. In *Advances in Neural Information Processing Systems 37*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; pp. 103031–103063.
17. Chen, H.; Song, J.; Han, C.; Xia, J.; Yokoya, N. ChangeMamba: Remote sensing change detection with spatiotemporal state space model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4409720.
18. Zhu, Q.; Li, H.; He, L.; Fan, L. SwinMamba: A hybrid local-global mamba framework for enhancing semantic segmentation of remotely sensed images. *arXiv* **2025**, arXiv:2509.20918.
19. Tang, Y.; Dong, P.; Tang, Z.; Chu, X.; Liang, J. Vmrnn: Integrating vision mamba and lstm for efficient and accurate spatiotemporal forecasting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 5663–5673.
20. Shi, Z.; Geng, H.; Wu, F.; Geng, L.; Zhuang, X. Radar-SR3: A Weather Radar Image Super-Resolution Generation Model Based on SR3. *Atmosphere* **2023**, *15*, 40.
21. Peebles, W.; Xie, S. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 4195–4205.
22. Yan, W.; Zhang, Y.; Abbeel, P.; Srinivas, A. Videogpt: Video generation using vq-vae and transformers. *arXiv* **2021**, arXiv:2104.10157.
23. Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; Qiao, Y. Videomamba: State space model for efficient video understanding. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2024; pp. 237–255.
24. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
25. Chen, L.; Cao, Y.; Ma, L.; Zhang, J. A deep learning-based methodology for precipitation nowcasting with radar. *Earth Space Sci.* **2020**, *7*, e2019EA000812.
26. Liu, Z.; Zhao, C.; Iandola, F.; Lai, C.; Tian, Y.; Fedorov, I.; Xiong, Y.; Chang, E.; Shi, Y.; Krishnamoorthi, R.; et al. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. In Proceedings of the Forty-First International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
27. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling laws for neural language models. *arXiv* **2020**, arXiv:2001.08361.
28. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
29. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
30. Veillette, M.; Samsi, S.; Mattioli, C. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In *Advances in Neural Information Processing Systems 33*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; pp. 22009–22019.

31. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. https://doi.org/10.1109/TIP.2003.819861.
32. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.