WaveSFNet: A Wavelet-Based Codec and Spatial–Frequency Dual-Domain Gating Network for Spatiotemporal Prediction

Xinyong Cai[1], Runming Xie[1], Hu Chen[2,3], Yuankai Wu[2,3,*]
[1]College of Software Engineering, Sichuan University [2]College of Computer Science, Sichuan University
[3]Sichuan University National Key Laboratory of Fundamental Science on Synthetic Vision
fenghuajuedai09@gmail.com, A372707325@126.com, huchen@scu.edu.cn, wuyk0@scu.edu.cn

***Abstract*—Spatiotemporal predictive learning aims to forecast future frames from historical observations in an unsupervised manner, and is critical to a wide range of applications. The key challenge is to model long-range dynamics while preserving high-frequency details for sharp multi-step predictions. Existing efficient recurrent-free frameworks typically rely on strided convolutions or pooling for sampling, which tends to discard textures and boundaries, while purely spatial operators often struggle to balance local interactions with global propagation. To address these issues, we propose WaveSFNet, an efficient framework that unifies a wavelet-based codec with a spatial–frequency dual-domain gated spatiotemporal translator. The wavelet-based codec preserves high-frequency subband cues during downsampling and reconstruction. Meanwhile, the translator first injects adjacent-frame differences to explicitly enhance dynamic information, and then performs dual-domain gated fusion between large-kernel spatial local modeling and frequency-domain global modulation, together with gated channel interaction for cross-channel feature exchange. Extensive experiments demonstrate that WaveSFNet achieves competitive prediction accuracy on Moving MNIST, TaxiBJ, and WeatherBench, while maintaining low computational complexity. Our code is available at https://github.com/fhjdqaq/WaveSFNet.**

***Index Terms*—spatiotemporal prediction, wavelet transform, frequency-domain learning, dual-domain gating, unsupervised learning**

I. Introduction

Spatiotemporal predictive learning aims to generate future sequences based on historical observations and is pivotal for applications ranging from weather forecasting [1], [2] and traffic flow prediction [3]–[5] to autonomous driving [6], [7]. A core challenge lies in the complex entanglement of multi-scale dynamics, where slowly evolving global trends coexist with abrupt local transients. Consequently, an effective model must capture long-range dependencies across the spatiotemporal field while rigorously preserving high-frequency details such as edges and textures to prevent error accumulation and blurring during multi-step prediction.

Existing approaches predominantly fall into recurrent-based and recurrent-free paradigms. Recurrent-based models, represented by ConvLSTM [1] and the PredRNN family [8]–[10], model temporal evolution via recursive state updates. Their step-by-step execution limits parallelization, resulting in high computational costs for high-resolution data. Conversely, recurrent-free methods [11], [12] formulate prediction as an end-to-end sequence-to-sequence mapping. These models typically employ an encoder–translator–decoder pipeline, offering superior parallel efficiency and scalability. However, most recurrent-free architectures rely on standard strided convolutions or pooling for spatial sampling. Without explicit constraints on frequency preservation, this process often causes high-frequency information loss, manifesting as blurred textures and degraded details, particularly in dynamic scenes.

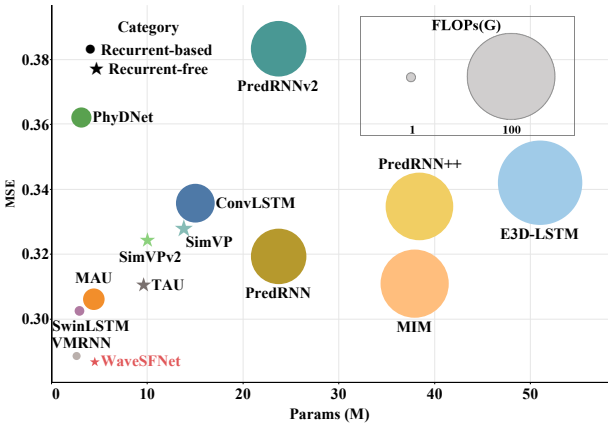

Fig. 1. Performance comparison on the TaxiBJ dataset. Bubble size denotes FLOPs. WaveSFNet achieves the lowest MSE with reduced complexity.

From the perspective of representation learning, purely spatial operators often struggle to balance global propagation with local detail retention [13], [14]. Frequency-domain modeling enables efficient global modulation for capturing long-range dependencies [15], [16]. Thus, a dual-domain strategy that synergizes spatial locality with spectral globality becomes essential. Complementing this, wavelet transforms provide a mathematically invertible framework that minimizes information loss during sampling by explicitly preserving high-frequency subbands [17], [18]. Therefore, it is promising to integrate dual-domain context modeling and multi-scale detail preservation within an efficient recurrent-free paradigm.

Motivated by these insights, we propose WaveSFNet, a wavelet-based spatial–frequency dual-domain gating network for spatiotemporal prediction. WaveSFNet adopts a streamlined encoder–translator–decoder architecture. The encoder

*Corresponding author.

leverages discrete wavelet transform for multi-scale downsampling and subband fusion, generating compact representations while preserving high-frequency cues. The translator stacks ST Blocks performing spatial–frequency dual-domain context extraction, fused via gated channel interaction. Finally, the decoder reconstructs the sequence through inverse discrete wavelet transform and a shallow skip connection. As highlighted in Fig. 1, this design achieves a superior trade-off between precision and efficiency on the TaxiBJ dataset, outperforming state-of-the-art baselines without relying on heavy attention mechanisms or recurrent inference.

Our main contributions are as follows:

- We propose WaveSFNet, which replaces conventional convolutional encoders and decoders with wavelet-based codecs to better preserve structural details in spatiotemporal prediction.
- We design a dual-domain gated spatiotemporal translator, which simultaneously models local dynamics and global propagation in the latent space via adaptive channel-wise fusion.
- Extensive experiments across diverse domains demonstrate that our method achieves state-of-the-art performance with a favorable balance between prediction quality and computational cost.

II. Related Work

A. Spatiotemporal Predictive Learning

Existing spatiotemporal predictive learning methods are commonly grouped into recurrent-based and recurrent-free architectures.

Early recurrent approaches mainly combine recurrent neural networks with convolutional operators. ConvLSTM [1] introduces convolutions into gated recurrent units, laying the foundation for spatiotemporal modeling. The PredRNN [8] family strengthens spatiotemporal memory to better capture short-term dynamics and reduce error accumulation. Later, MIM [19] enhances change modeling via differential memory, while PhyDNet [20] incorporates physical priors to disentangle dynamics. More recently, SwinLSTM [21] adopts window-based attention to expand spatial dependency modeling, and VMRNN [22] introduces state-space modeling to improve long-sequence efficiency. Despite their strengths, recurrent models rely on step-by-step updates, limiting parallelism and often increasing training and inference costs.

Recurrent-free architectures remove the sequential bottleneck by formulating prediction as an end-to-end sequence-to-sequence mapping. SimVP [11] establishes a strong baseline with a concise encoder–translator–decoder pipeline, where the translator can be replaced by MetaFormer-style [23], [24] modular backbones to compare interaction mechanisms. TAU [12] decomposes temporal modeling for parallel motion capture. Earthformer [25] targets high-resolution Earth system forecasting with scalable cuboid attention. MMVP [26] adopts a dual-stream design to separate motion and appearance. TAT [27] introduces triplet attention that factorizes interactions over temporal, spatial, and channel dimensions. While recurrent-free methods are highly parallelizable, purely spatial designs or fixed receptive fields may degrade details in highly dynamic scenes, especially high-frequency content.

B. Wavelets and Frequency-Domain Learning

Spectral methods use Fourier transforms for efficient global frequency modulation. GFNet [15] introduces learnable frequency-domain filters as an efficient alternative to expensive attention. FourCastNet [28] uses adaptive Fourier neural operators for global mixing in Earth system prediction. PastNet [29] incorporates frequency priors into spatiotemporal forecasting, and SpectFormer [30] combines spectral gating with attention to balance global modulation and local interactions. Standard Fourier representations emphasize global spectral modeling, and the incorporation of spatial locality can be considered for non-stationary signals with strong transients.

Wavelet-based approaches exploit multi-resolution analysis with spatial–frequency locality, enabling detail-preserving sampling. MWCNN [17] replaces conventional sampling with wavelet transforms to enlarge the receptive field while reducing information loss. WaveViT [31] integrates wavelet-based sampling into vision transformers for unified multi-scale learning. For spatiotemporal prediction, WaST [32] leverages multi-scale subbands to retain fine-grained structures during temporal forecasting. Compared with standard sampling schemes, wavelet-based models preserve local details across scales and provide a faithful basis for subsequent modeling.

III. Method

A. Overall Framework

Given an input spatiotemporal sequence $\mathbf{X} \in \mathbb{R}^{B\times T_{\text{in}}\times C\times H\times W}$, we learn a parametric mapping f_θ to model the spatiotemporal dynamics and predict future frames:

$$\widehat{\mathbf{Y}} = f_\theta(\mathbf{X}) \in \mathbb{R}^{B\times T_{\text{out}}\times C\times H\times W}. \tag{1}$$

As depicted in Fig. 2, the overall framework consists of three components: (1) **Wavelet-based multi-scale encoder**, which decomposes each input frame into low-frequency structures and high-frequency details and fuses them into compact latent representations; (2) **Spatiotemporal translator**, which injects adjacent-frame difference cues into the coarse-scale latent space and extracts contextual information in parallel from spatial and frequency domains, followed by gated channel interaction; (3) **Wavelet-symmetric decoder**, which progressively reconstructs spatial resolution via inverse wavelet transforms and produces the predicted frames.

B. Wavelet-based Multi-scale Encoding

For each time step $t = 1, \ldots, T_{\text{in}}$, the input frame is denoted as $\mathbf{X}_t \in \mathbb{R}^{C\times H\times W}$. Shallow features are first extracted:

$$\mathbf{F}_t^{(0)} = \text{Conv}_{3\times3}\Big(\text{GELU}(\text{Conv}_{3\times3}(\mathbf{X}_t))\Big). \tag{2}$$

The resulting shallow features satisfy $\mathbf{F}_t^{(0)} \in \mathbb{R}^{C_0\times H\times W}$, and are retained for high-resolution detail compensation in decoding.

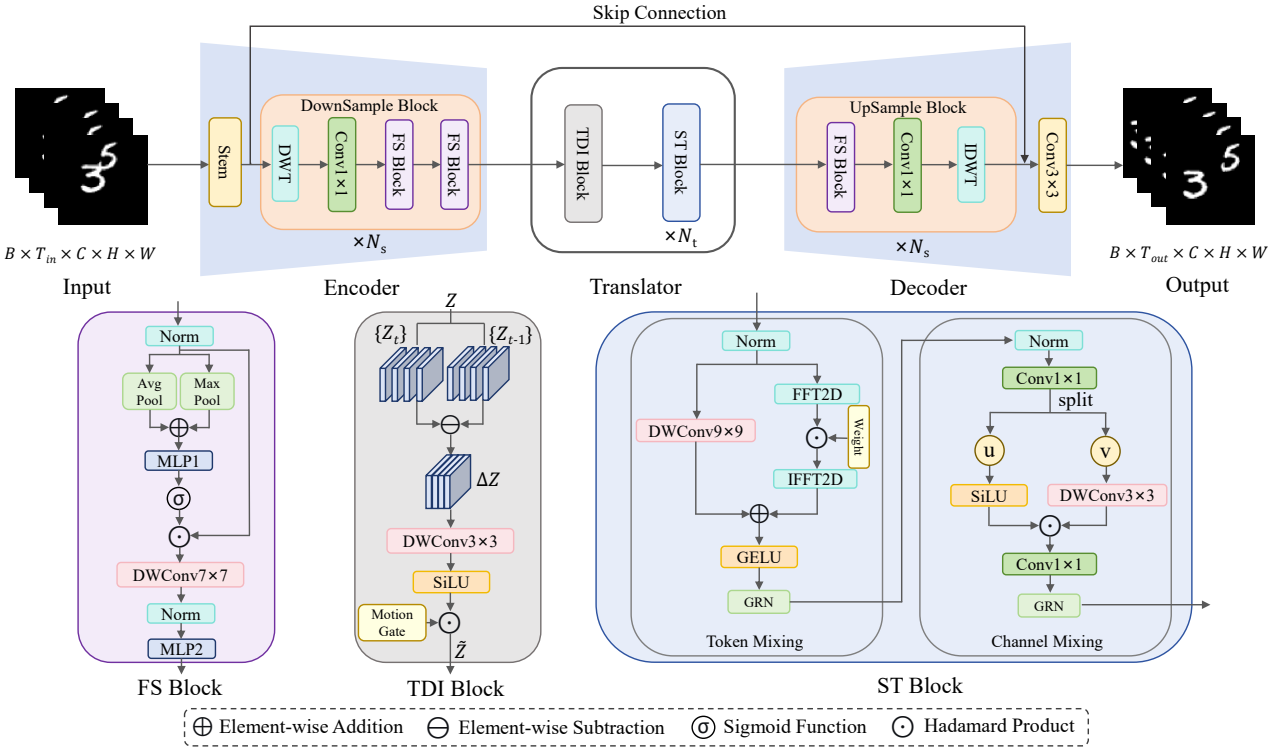

Fig. 2. **Overall architecture and core modules of WaveSFNet.** WaveSFNet follows an encoder–translator–decoder design. A wavelet-based multi-scale encoder extracts latent features from input frames. A **TDI Block** injects adjacent-frame differences, and N_t stacked **ST Blocks** apply spatial–frequency dual-domain gating after packing time into channels. A wavelet-symmetric decoder reconstructs predictions.

Subsequently, N_s wavelet downsampling stages are applied. We employ the **Haar** wavelet for 2D DWT/IDWT, which is a **parameter-free and invertible** linear transform, enabling stable multi-scale decomposition and reconstruction. At the i-th stage, a 2D discrete wavelet transform decomposes the features into four subbands:

$$\left(\mathbf{F}_t^{(i),LL}, \mathbf{F}_t^{(i),HL}, \mathbf{F}_t^{(i),LH}, \mathbf{F}_t^{(i),HH}\right) = \mathcal{W}\left(\mathbf{F}_t^{(i-1)}\right). \quad (3)$$

The four subbands are concatenated along the channel dimension and projected via a pointwise convolution:

$$\mathbf{E}_t^{(i)} = \mathrm{Conv}_{1\times1}\left(\mathrm{Concat}_c\left(\{\mathbf{F}_t^{(i),b}\}_{b\in\mathcal{B}}\right)\right), \quad (4)$$

where $\mathbf{E}_t^{(i)} \in \mathbb{R}^{C_s\times H_i\times W_i}$, $\mathcal{B} = \{LL, HL, LH, HH\}$ and $H_i = H/2^i$, $W_i = W/2^i$.

This fusion compresses complementary subband information into a unified channel space, forming compact multi-scale representations.

Frequency-selective enhancement is then applied to $\mathbf{E}_t^{(i)}$ using two FS Blocks at each scale to obtain $\mathbf{F}_t^{(i)}$ for stable representation learning. Specifically, normalized features are gated by global average and max pooled statistics, refined by large-kernel depthwise convolution, and mixed by pointwise channel interaction.

C. Temporal Difference Injection

To enhance sensitivity to rapid temporal variations, we inject adjacent-frame differences in the coarse latent space via a TDI (Temporal Difference Injection) Block.

We first project the last-scale features $\mathbf{F}_t^{(N_s)}$ into compact latent representations $\mathbf{Z}_t \in \mathbb{R}^{C_z\times H_{N_s}\times W_{N_s}}$ via a frame-wise pointwise convolution, and compute adjacent-frame differences $\Delta\mathbf{Z}_t$ (with $\Delta\mathbf{Z}_1 = \mathbf{0}$). We then inject these differences by extracting local motion cues using a depthwise convolution and modulating them with channel-wise gating:

$$\widetilde{\mathbf{Z}}_t = \mathbf{Z}_t + \mathbf{g} \odot \mathrm{SiLU}\Big(\mathrm{DWConv}_{3\times3}(\Delta\mathbf{Z}_t)\Big), \quad (5)$$

where $\mathbf{g}$ denotes learnable gating coefficients broadcastable over spatial dimensions. This operation introduces explicit short-term dynamics in the coarse latent space, enhancing sensitivity to rapid variations.

D. Dual-domain Gated Spatiotemporal Translation

The injected sequence $\widetilde{\mathbf{Z}} = \{\widetilde{\mathbf{Z}}_t\}_{t=1}^{T_{\mathrm{in}}}$ is fed into a spatiotemporal translator composed of N_t stacked ST Blocks. We stack the temporal dimension into channels and denote the resulting channel-stacked features as $\widetilde{\mathbf{Z}}^{\mathrm{p}} \in \mathbb{R}^{C_t\times H_{N_s}\times W_{N_s}}$, where $C_t = T_{\mathrm{in}}C_z$. Each ST Block consists of dual-domain context extraction and gated channel interaction. The spatial branch emphasizes local neighborhood structures, while the frequency branch captures global patterns and long-range dependencies, and the two complement each other.

1) Dual-domain Context Extraction: The spatial branch employs a 9×9 depthwise convolution:

$$\mathbf{S} = \mathrm{DWConv}_{9\times9}(\widetilde{\mathbf{Z}}^{\mathrm{p}}). \quad (6)$$

The frequency branch applies a 2D real Fourier transform, modulates the spectrum with learnable complex-valued weights, and transforms back to the spatial domain:

$$\mathbf{P} = \mathcal{F}^{-1}\big(\mathcal{F}(\widetilde{\mathbf{Z}}^{\mathrm{p}}) \odot \mathbf{\Psi}\big), \tag{7}$$

where $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote the 2D real Fourier transform and its inverse, respectively, and $\mathbf{\Psi}$ represents learnable complex weights. As the frequency weights operate globally over the spectrum, this branch provides modulation with a global receptive field, which is well suited for modeling large-scale structure propagation and long-range dependencies.

We fuse the two branches as $\mathbf{U} = \mathbf{S} + \mathbf{P}$, and implement it with a residual calibration for stable training.

2) Gated Channel Interaction: Gated channel interaction is applied to enable adaptive cross-channel information exchange. We first use a pointwise convolution to expand the contextual features, and then split the result along the channel dimension into two parts, $\mathbf{U}^{(u)}$ and $\mathbf{U}^{(v)}$, where $\mathbf{U}^{(u)}$ provides gating signals and $\mathbf{U}^{(v)}$ carries feature content.

A 3×3 depthwise convolution is applied to the feature branch, followed by element-wise modulation with a SiLU gate and projection back to the original channel dimension:

$$\mathbf{G} = \mathrm{Conv}_{1\times1}\Big(\mathrm{SiLU}(\mathbf{U}^{(u)}) \odot \mathrm{DWConv}_{3\times3}(\mathbf{U}^{(v)})\Big). \tag{8}$$

This gated interaction enables nonlinear channel mixing, facilitating the exchange of frequency-aware semantics and temporal cues across channels. GRN [33], layer scale [34], and DropPath are applied throughout the spatiotemporal translator to stabilize training and provide regularization. After stacking N_t ST Blocks, we reshape the output to unstack time from channels, followed by a frame-wise pointwise convolution that maps C_z back to C_s, producing $\{\mathbf{Q}_t\}_{t=1}^{T_{\mathrm{in}}} \in \mathbb{R}^{T_{\mathrm{in}}\times C_s\times H_{N_s}\times W_{N_s}}$.

E. Wavelet-symmetric Decoding

The decoder takes the translated latent sequence $\{\mathbf{Q}_t\}_{t=1}^{T_{\mathrm{in}}}$ as input and mirrors the encoder with N_s upsampling stages.

At each stage, the latent features are refined by an FS Block to better support inverse wavelet reconstruction, followed by a pointwise convolution and a 2D inverse wavelet transform to restore the spatial resolution from (H_i, W_i) to (H_{i-1}, W_{i-1}).

After the final stage, we fuse the shallow features $\mathbf{F}_t^{(0)}$ from the encoder via a skip connection to recover fine-grained details. A 3×3 readout convolution is then applied to produce the predicted frames $\widehat{\mathbf{Y}}_t \in \mathbb{R}^{C\times H\times W}$, $t = 1, \ldots, T_{\mathrm{out}}$, and the outputs are stacked over time to form $\widehat{\mathbf{Y}}$.

By default, $T_{\mathrm{out}} = T_{\mathrm{in}}$. When $T_{\mathrm{out}} < T_{\mathrm{in}}$, the first T_{out} frames of the predicted sequence are retained; when $T_{\mathrm{out}} > T_{\mathrm{in}}$, an autoregressive rollout strategy is adopted, where the previously predicted sequence is iteratively fed back as input until reaching length T_{out}.

IV. Experiments

A. Experimental Setup

1) Datasets: Following the OpenSTL benchmark protocol [24], we evaluate our method on three datasets from distinct domains.

Moving MNIST [35] is a synthetic motion benchmark of two moving handwritten digits on 64×64 grids. We use 10,000 sequences for training and 10,000 for testing. **TaxiBJ** [3] contains real-world Beijing traffic inflow and outflow fields on 32×32 grids with two channels. We use 20,461 samples for training and 500 for testing. **WeatherBench** [36] is a global weather forecasting benchmark with multiple spatial resolutions. We evaluate two settings on the 5.625° version. (1) *Single-variable* forecasting for temperature (T2M), humidity (R), wind components (UV10), and cloud cover (TCC), using data from 2010–2015 for training, 2016 for validation, and 2017–2018 for testing. (2) *Multi-variable (MV)* forecasting jointly predicts humidity (R), temperature (T), longitude wind (U), and latitude wind (V) at 150, 500, and 850 hPa, using a longer training period (1979–2015) and the same validation (2016) and testing (2017–2018) years.

2) Metrics: We employ standard metrics for comprehensive evaluation. MSE and MAE are used to quantify prediction errors (lower is better), with RMSE additionally reported for weather tasks. SSIM and PSNR are used to assess visual quality and structural fidelity (higher is better). Furthermore, model efficiency is evaluated by the number of parameters (Params) and floating-point operations (FLOPs).

3) Implementation Details: Our model is implemented using PyTorch within the OpenSTL framework [24]. We utilize the Adam optimizer for all experiments, minimizing the Mean Squared Error (MSE) loss function. The learning rate is scheduled via a OneCycle strategy for Moving MNIST and a Cosine Annealing strategy for the other datasets. The

TABLE I
DETAILED EXPERIMENTAL CONFIGURATIONS FOR DIFFERENT DATASETS. C_s DENOTES THE BASE CHANNEL WIDTH OF THE WAVELET CODEC, AND C_t DENOTES THE TIME-PACKED CHANNEL WIDTH USED BY ST BLOCKS.

Dataset		N_s	N_t	C_s	C_t	(C, H, W)	Batch Size	T_{in}	T_{out}	LR	Epochs
Moving MNIST		2	8	64	720	$(1, 64, 64)$	16	10	10	9.2×10^{-4}	2000
TaxiBJ		1	8	32	192	$(2, 32, 32)$	16	4	4	1.5×10^{-3}	50
WeatherBench	T2M	1	8	32	264	$(1, 32, 64)$	16	12	12	2×10^{-3}	50
	TCC	1	8	32	264	$(1, 32, 64)$	16	12	12	4×10^{-3}	50
	UV10	1	8	32	264	$(2, 32, 64)$	16	12	12	2×10^{-3}	50
	R	1	8	32	264	$(1, 32, 64)$	16	12	12	3×10^{-3}	50
	MV	1	8	32	264	$(12, 32, 64)$	16	4	4	3×10^{-3}	50

specific architectural hyperparameters and detailed training configurations are summarized in Table I.

B. Comparisons with State-of-the-Art

We compare our method with two categories: *recurrent-based* approaches built on recurrent networks (e.g., ConvLSTM, PredRNN, and VMRNN), and *recurrent-free* approaches without recurrence (e.g., SimVP, TAU, and MogaNet).

TABLE II
QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON MOVING MNIST.

Category	Method	MSE ↓	SSIM ↑
Recurrent-based	ConvLSTM [1]	103.3	0.707
	FRNN [37]	69.7	0.813
	PredRNN [8]	56.8	0.867
	MIM [19]	44.2	0.910
	MAU [38]	27.6	0.937
	PhyDNet [20]	24.4	0.947
	CrevNet [39]	22.3	0.949
	SwinLSTM [21]	17.7	0.962
	VMRNN [22]	16.5	0.965
Recurrent-free	SimVP [11]	23.8	0.948
	MMVP [26]	22.2	0.952
	TAU [12]	19.8	0.957
	TAT [27]	17.6	0.960
	Ours	**15.8**	**0.966**

Results on Synthetic Motion. Table II reports quantitative results on Moving MNIST. Our method achieves the best MSE (15.8) and SSIM (0.966). This superior structural fidelity stems from our wavelet-based codec explicitly preserving high-frequency edges, which effectively mitigates the texture blurring typically observed in long-horizon predictions.

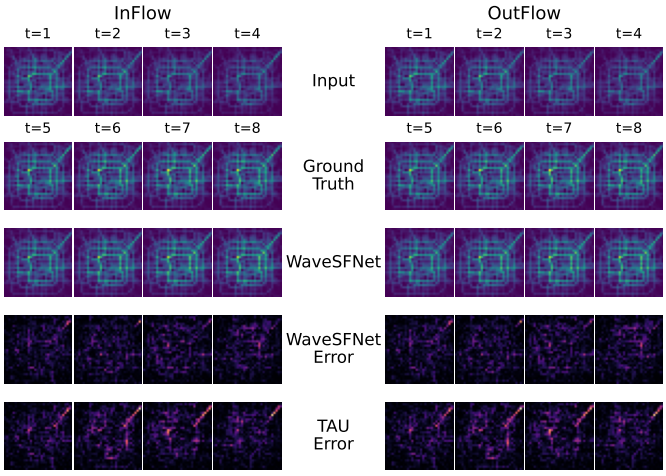

Fig. 3. Qualitative visualizations of WaveSFNet on TaxiBJ.

Results on Urban Traffic Flow. Urban traffic flow forecasting requires capturing both nonlinear spatiotemporal dependencies and periodic variations. As shown in Table III, our method consistently leads in all metrics. Qualitative comparisons in Fig. 3 further validate this superiority; the error heatmaps demonstrate that WaveSFNet yields significantly lower residuals than TAU, indicating a more precise capture of complex flow dynamics. Crucially, we achieve this with highly competitive computational cost (4.5M Params, 1.1G FLOPs).

TABLE III
QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON TAXIBJ.

Category	Method	Params (M)	FLOPs (G)	MSE↓	MAE↓	SSIM↑	PSNR↑
Recurrent-based	ConvLSTM [1]	15.0	20.7	0.3358	15.32	0.9836	39.45
	PredRNN [8]	23.7	42.4	0.3194	15.31	0.9838	39.51
	PredRNN++ [9]	38.4	63.0	0.3348	15.37	0.9834	39.47
	E3D-LSTM [40]	51.0	98.19	0.3421	14.98	0.9842	39.64
	PhyDNet [20]	3.1	5.6	0.3622	15.53	0.9828	39.46
	MIM [19]	37.9	64.1	0.3110	14.96	0.9847	39.65
	MAU [38]	4.4	6.4	0.3062	15.26	0.9840	39.52
	PredRNNv2 [10]	23.7	42.6	0.3834	15.55	0.9826	39.49
	SwinLSTM [21]	2.9	1.3	0.3026	15.00	0.9843	–
	VMRNN [22]	**2.6**	**0.9**	0.2887	14.69	0.9858	–
Recurrent-free	SimVP [11]	13.8	3.6	0.3282	15.45	0.9835	39.45
	SimVPv2 [41]	10.0	2.6	0.3246	15.03	0.9844	39.71
	TAU [12]	9.6	2.5	0.3108	14.93	0.9848	39.74
	MogaNet [42]	10.0	2.6	0.3114	15.06	0.9847	39.70
	Ours	4.5	1.1	**0.2870**	**14.68**	**0.9859**	**39.88**

This optimal accuracy–efficiency trade-off makes our method suitable for deployment on resource-constrained edge devices in intelligent transportation systems.

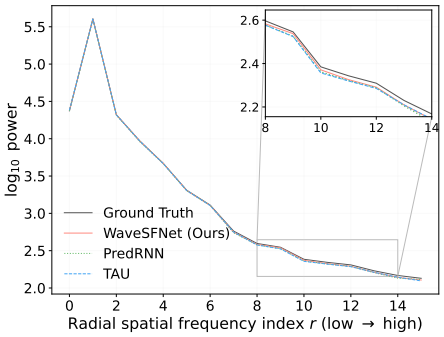

Fig. 4. Frequency spectrum analysis on the WeatherBench T2M dataset. The plot displays the Radially Averaged Power Spectral Density, where the x-axis represents the radial spatial frequency and the y-axis denotes the log power spectrum. The inset provides a zoomed-in view of the high-frequency region.

Results on Weather Forecasting. Weather forecasting is challenging due to chaotic atmospheric dynamics and complex multi-scale dependencies. For single-variable forecasting (Table IV), our method remains robust across diverse variables. This trend aligns with our motivation: weather fields exhibit smooth structures and high-frequency variations, where dual-domain modeling can be beneficial.

To assess structural fidelity, we conduct a spectral analysis on the T2M dataset by computing the Radially Averaged Power Spectral Density (RAPSD) on the last predicted frame of all 17,495 test sequences. As shown in Fig. 4, accurately capturing high-frequency components remains a formidable challenge, with all models deviating from the Ground Truth. PredRNN and TAU exhibit a more pronounced high-frequency power decay, while WaveSFNet demonstrates a closer spectral alignment, suggesting that our overall architecture better mitigates the smoothing bias.

For multi-variable forecasting (Table V), our method achieves the best overall accuracy with only 3.73G FLOPs, suggesting that WaveSFNet can capture cross-variable dependencies under limited computational budgets, scaling elegantly across diverse meteorological factors.

TABLE IV
QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON WEATHERBENCH SINGLE-VARIABLE TASKS, INCLUDING UV10, T2M, TCC, AND R.

Method	Params (M)	FLOPs (G)	UV10			T2M			TCC			R		
			MSE↓	MAE↓	RMSE↓	MSE↓	MAE↓	RMSE↓	MSE↓	MAE↓	RMSE↓	MSE↓	MAE↓	RMSE↓
ConvLSTM [1]	14.98	136	1.8976	0.9215	1.3775	1.5210	0.7949	1.2330	0.0494	0.1542	0.2223	35.1460	4.0120	5.9280
PredRNN [8]	23.57	278	1.8810	0.9068	1.3715	1.3310	0.7246	1.1540	0.0550	0.1588	0.2346	37.6110	4.0960	6.1330
PredRNN++ [9]	38.40	413	1.8727	0.9019	1.3685	1.4580	0.7676	1.2070	0.0548	0.1544	0.2341	45.9930	4.7310	6.7820
MAU [38]	5.46	39.6	1.9001	0.9194	1.3784	1.2510	0.7036	1.1190	0.0496	0.1516	0.2226	34.5290	4.0040	5.8760
SimVP [11]	14.67	8.03	1.9993	0.9510	1.4140	1.2380	0.7037	1.1130	0.0477	0.1503	0.2183	34.3550	3.9940	5.8610
HorNet [43]	12.42	6.85	1.5539	0.8254	1.2466	1.2010	0.6906	1.0960	0.0469	0.1475	0.2166	32.0810	3.8260	5.6640
TAU [12]	12.22	6.70	1.5925	0.8426	1.2619	1.1620	0.6707	1.0780	0.0472	0.1460	0.2173	31.8310	3.8180	5.6420
MogaNet [42]	12.76	7.01	1.6072	0.8451	1.2678	1.1520	0.6665	1.0730	0.0470	0.1480	0.2168	31.7950	3.8160	5.6390
WaST [32]	**3.90**	**1.50**	-	-	-	1.0980	0.6338	1.0440	-	**0.1452**	0.2150	-	**3.6940**	5.5690
Ours	8.73	4.06	**1.4601**	**0.8063**	**1.2084**	**1.0458**	**0.6311**	**1.0226**	**0.0459**	0.1459	**0.2143**	**30.4813**	3.7363	**5.5210**

TABLE V
QUANTITATIVE COMPARISON OF DIFFERENT METHODS ON WEATHERBENCH MULTI-VARIABLE FORECASTING (R, T, U, AND V). ALL METRICS ARE AVERAGED OVER THE FOUR VARIABLES.

Method	Params (M)	FLOPs (G)	MSE↓	MAE↓	RMSE↓
ConvLSTM [1]	15.50	43.33	108.81	5.7439	8.1810
PredRNN [8]	24.56	88.02	104.16	5.5373	7.9553
PredRNN++ [9]	39.31	129.0	106.77	5.5821	8.0568
MIM [19]	41.71	35.77	121.95	6.2786	8.7376
MAU [38]	5.46	12.07	106.13	5.6487	7.9928
PredRNNv2 [10]	24.58	88.49	108.94	5.7747	8.1872
SimVP [11]	13.80	7.26	108.50	5.7360	8.1165
SimVPv2 [41]	9.96	5.25	103.36	5.4856	7.9059
ConvMixer [44]	**0.85**	**0.49**	112.76	5.9114	8.3238
HorNet [43]	9.68	5.12	104.21	5.5181	7.9854
MogaNet [42]	9.97	5.25	98.664	5.3003	7.6539
TAU [12]	9.55	5.01	99.428	5.3282	7.6855
Ours	8.26	3.73	**96.141**	**5.2550**	**7.5569**

C. Ablation Study

We perform ablation studies under the standard protocol on TaxiBJ and WeatherBench (TCC). All variants share the same training pipeline; only the specified component is modified.

TABLE VI
ABLATION RESULTS ON TAXIBJ AND TCC.

Variant	TaxiBJ			TCC	
	MSE↓	MAE↓	SSIM↑	MSE↓	MAE↓
Spatial-only	0.2949	14.76	0.9854	0.04604	0.1494
Frequency-only	0.3014	14.72	0.9850	0.04653	0.1518
w/o TDI Block	0.2962	14.76	0.9854	0.04598	0.1489
Conv codec	0.3198	14.92	0.9848	0.04609	0.1472
MLP mixer	0.2997	14.91	0.9850	0.04620	0.1476
WaveSFNet	**0.2870**	**14.68**	**0.9859**	**0.04593**	**0.1459**

As summarized in Table VI, we evaluate five variants: *Spatial-only* removes the frequency branch in the dual-domain context extraction; *Frequency-only* removes the spatial branch; *w/o TDI Block* removes the TDI Block; *Conv codec* replaces the wavelet-based codec with a convolutional one; *MLP mixer* replaces gated channel interaction with standard MLP mixing.

Table VI shows that the full model achieves the best overall performance across both datasets. Removing either branch degrades accuracy, highlighting the complementarity between local spatial modeling and global spectral modulation. The performance drop observed in the *w/o TDI Block* variant suggests that explicitly injecting short-term motion cues enhances the model's sensitivity to rapid temporal dynamics. Replacing the wavelet codec leads to the largest drop, suggesting that wavelet-based multi-scale decomposition mitigates information loss during sampling and supports more faithful reconstruction. Replacing multiplicative gating with MLP mixing also harms performance, indicating the benefit of gated channel interaction for feature exchange.

V. CONCLUSION

In this work, we propose WaveSFNet, an efficient recurrent-free spatiotemporal prediction architecture. By integrating a wavelet codec with a spatial–frequency dual-domain gating mechanism, WaveSFNet effectively alleviates high-frequency information loss during sampling while jointly modeling local dynamic variations and global long-range dependencies. Extensive experiments on synthetic motion, urban traffic flow, and global weather forecasting benchmarks demonstrate that WaveSFNet achieves state-of-the-art performance with a favorable balance between accuracy and computational cost. These results validate the effectiveness of incorporating multi-resolution spectral priors for high-fidelity spatiotemporal modeling, and suggest a promising direction for learning complex dynamic systems.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 62406206), the Fundamental Research Funds for the Central Universities, and the Sichuan Provincial Natural Science Foundation (No. 2026NSFSC0426). We also appreciate the foundational work of pioneers in this field.

REFERENCES

[1] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[2] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and F. Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[3] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017, pp. 1655–1661.

[4] S. Fang, Q. Zhang, G. Meng, S. Xiang, and C. Pan, "Gstnet: Global spatial-temporal network for traffic flow prediction." in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 2286–2293.

[5] J. Cheng, K. Li, Y. Liang, L. Sun, J. Yan, and Y. Wu, "Rethinking urban mobility prediction: A multivariate time series forecasting approach," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[6] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4194–4202.

[7] Y.-H. Kwon and M.-G. Park, "Predicting future frames using retrospective cycle gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1811–1820.

[8] Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu, "Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[9] Y. Wang, Z. Gao, M. Long, J. Wang, and P. S. Yu, "Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5123–5132.

[10] Y. Wang, H. Wu, J. Zhang, Z. Gao, J. Wang, P. S. Yu, and M. Long, "Predrnn: A recurrent neural network for spatiotemporal predictive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2208–2225, 2022.

[11] Z. Gao, C. Tan, L. Wu, and S. Z. Li, "Simvp: Simpler yet better video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3170–3180.

[12] C. Tan, Z. Gao, L. Wu, Y. Xu, J. Xia, S. Li, and S. Z. Li, "Temporal attention unit: Towards efficient spatiotemporal predictive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 770–18 782.

[13] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[15] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," *Advances in Neural Information Processing Systems*, vol. 34, pp. 980–993, 2021.

[16] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "Fnet: Mixing tokens with fourier transforms," in *Proceedings of the 2022 Conference of the north American chapter of the Association for Computational Linguistics: human language technologies*, 2022, pp. 4296–4313.

[17] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-cnn for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 773–782.

[18] Q. Li, L. Shen, S. Guo, and Z. Lai, "Wavelet integrated cnns for noise-robust image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7245–7254.

[19] Y. Wang, J. Zhang, H. Zhu, M. Long, J. Wang, and P. S. Yu, "Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9154–9162.

[20] V. L. Guen and N. Thome, "Disentangling physical dynamics from unknown factors for unsupervised video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 474–11 484.

[21] S. Tang, C. Li, P. Zhang, and R. Tang, "Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 470–13 479.

[22] Y. Tang, P. Dong, Z. Tang, X. Chu, and J. Liang, "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 Workshops*, 2024, pp. 5663–5673.

[23] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 819–10 829.

[24] C. Tan, S. Li, Z. Gao, W. Guan, Z. Wang, Z. Liu, L. Wu, and S. Z. Li, "Openstl: A comprehensive benchmark of spatio-temporal predictive learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 69 819–69 831, 2023.

[25] Z. Gao, X. Shi, H. Wang, Y. Zhu, Y. B. Wang, M. Li, and D.-Y. Yeung, "Earthformer: Exploring space-time transformers for earth system forecasting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 390–25 403, 2022.

[26] Y. Zhong, L. Liang, I. Zharkov, and U. Neumann, "Mmvp: Motion-matrix-based video prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4273–4283.

[27] X. Nie, X. Chen, H. Jin, Z. Zhu, Y. Yan, and D. Qi, "Triplet attention transformer for spatiotemporal predictive learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7036–7045.

[28] J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli *et al.*, "Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators," *arXiv preprint arXiv:2202.11214*, 2022.

[29] H. Wu, F. Xu, C. Chen, X.-S. Hua, X. Luo, and H. Wang, "Pastnet: Introducing physical inductive biases for spatio-temporal video prediction," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 2917–2926.

[30] B. N. Patro, V. P. Namboodiri, and V. S. Agneeswaran, "Spectformer: Frequency and attention is what you need in a vision transformer," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025, pp. 9525–9536.

[31] T. Yao, Y. Pan, Y. Li, C.-W. Ngo, and T. Mei, "Wave-vit: Unifying wavelet and transformers for visual representation learning," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 328–345.

[32] X. Nie, Y. Yan, S. Li, C. Tan, X. Chen, H. Jin, Z. Zhu, S. Z. Li, and D. Qi, "Wavelet-driven spatiotemporal predictive learning: Bridging frequency and time variations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4334–4342.

[33] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 133–16 142.

[34] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 32–42.

[35] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using lstms," in *International Conference on Machine Learning*. PMLR, 2015, pp. 843–852.

[36] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, "Weatherbench: a benchmark data set for data-driven weather forecasting," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 11, p. e2020MS002203, 2020.

[37] M. Oliu, J. Selva, and S. Escalera, "Folded recurrent neural networks for future video prediction," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 716–731.

[38] Z. Chang, X. Zhang, S. Wang, S. Ma, Y. Ye, X. Xinguang, and W. Gao, "Mau: A motion-aware unit for video prediction and beyond," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26 950–26 962, 2021.

[39] W. Yu, Y. Lu, S. Easterbrook, and S. Fidler, "Efficient and information-preserving future frame prediction and beyond," in *International Conference on Learning Representations*, 2020.

[40] Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, "Eidetic 3d lstm: A model for video prediction and beyond," in *International Conference on Learning Representations*, 2018.

[41] C. Tan, Z. Gao, S. Li, and S. Z. Li, "Simvpv2: Towards simple yet powerful spatiotemporal predictive learning," *IEEE Transactions on Multimedia*, 2025.

[42] S. Li, Z. Wang, Z. Liu, C. Tan, H. Lin, D. Wu, Z. Chen, J. Zheng, and S. Z. Li, "Moganet: Multi-order gated aggregation network," in *International Conference on Learning Representations*, 2024.

[43] Y. Rao, W. Zhao, Y. Tang, J. Zhou, S.-L. Lim, and J. Lu, "Hornet: Efficient high-order spatial interactions with recursive gated convolutions," *Advances in Neural Information Processing Systems*, 2022.

[44] A. Trockman and J. Z. Kolter, "Patches are all you need?" *Transactions on Machine Learning Research*, 2023, featured Certification.