Beyond RNNs and CNNs: A Unified Transformer for Spatiotemporal Prediction

Haowei Mei, Zhe Huang, Zhaoyu Qi, Bo Li*
Xi'an Jiaotong University, Xi'an, China *Corresponding author: boblee@xjtu.edu.cn

***Abstract*—Spatiotemporal prediction has long been dominated by recurrent and convolutional neural networks. While Transformers have achieved remarkable success in other domains, their application as pure, standalone architectures for this task remains notably under explored. Most existing approaches merely augment RNN or CNN backbones with attention mechanisms, preserving their inherent inductive biases. In this work, we present MeldFormer, a simple yet powerful model built entirely upon Transformers. It introduces a token grouping and fusion mechanism that unifies spatiotemporal modeling within a single, coherent attention-based framework, eliminating the need for separate modules and enabling efficient global context modeling. Extensive experiments on four standard benchmarks, including Moving MNIST, WeatherBench, TaxiBJ, and KTH, demonstrate that MeldFormer achieves state-of-the-art prediction accuracy while maintaining significantly lower computational complexity. Code is available at: https://github.com/May2001/MeldFormer.**

***Index Terms*—spatio-temporal prediction, computer vision**

I. Introduction

The future is fascinating, and predicting the future is even more so. Spatio-temporal prediction, which forecasts future spatio-temporal patterns based on historical observations, has attracted increasing attention. It has numerous practical applications, such as object motion prediction [1], precipitation forecasting [2]–[4], traffic flow prediction [5], [6], and human body motion prediction [7]. Now, research efforts have mainly concentrated on a crucial issue, that is, how to perceive the spatial motion information of objects over past time.

Early methods typically relied on recurrent neural networks to model temporal dependencies, but they encountered difficulties in long-term information retention and computational efficiency. Regarding the first issue, it is usually addressed by incorporating new information transmission paths into complex RNN architectures. For example, PredRNN++ [8] introduced the gradient highway unit to enhance the long-term information propagation across time steps. However, for the second issue, due to the inherent structure of RNNs, sequences can only be computed serially and cannot be parallelized, making it difficult to improve computational efficiency. SimVP [9] constructed a pure convolutional model, implementing a non-recursive architecture, which significantly improved computational efficiency. On this basis, variants such as TAU [10] and WaST [11] have emerged successively to capture the changes in the temporal dimension, achieving performance comparable to that of RNN models with more efficient structures. Nevertheless, convolutional neural networks, which are designed for extracting spatial features, have an inherent disadvantage in capturing temporal changes.

Recently, Transformer-based architectures have become an effective solution to these two problems. By leveraging the self-attention mechanism, they can effectively capture long-range temporal correlations, thus solving the problem of long-term information transmission. Additionally, they improve computational efficiency through parallel computing. However, pure transformer-based architectures remain insufficiently explored, primarily attributed to the high computational complexity incurred by sequence length. Mainstream approaches [12], [13] reduce this complexity by decoupling spatial and temporal attention. In contrast to these methods, our work compresses and fuses spatiotemporal information, achieving superior prediction performance.

In this paper, we propose MeldFormer, a spatio-temporal prediction learning architecture of Transformer that realizes the compression and fusion of tokens. The core idea is as follows: patches tokens that are at the same position but in different time steps are grouped into the same set. Then, the hidden dimensions of the features of tokens within the group are compressed. Finally, the tokens within the group are concatenated and fused into new tokens. Through this grouping approach, the attention calculation is divided into two forms: local and global. That is, the self-attention calculation within the group and the global self-attention calculation between groups after fusion.

The contributions of this paper are summarized as follows:

- We propose a method for token compression and fusion as well as decompression and decomposition, which significantly reduces the computational complexity while preserving spatio-temporal information.
- We have designed a new architecture for spatio-temporal prediction tasks. Through local self-attention calculation and global self-attention calculation, it can effectively model spatial and temporal dependencies.
- We conduct comprehensive evaluations of the proposed model on four diverse datasets. The experimental outcomes demonstrate that MeldFormer exhibits outstanding performance across all these datasets, highlighting its superior effectiveness and versatility in spatio-temporal prediction tasks.

II. Related Work

A. Decoupled Spatiotemporal Modeling

RNN and CNN are classical models for addressing spatiotemporal prediction problems. RNN models, leveraging

their internal recurrent structure to process information at each timestep, inherently decouple temporal and spatial dimensions, which aligns well with the sequential nature of temporal data. Since the introduction of the ConvLSTM network by Shi et al [2]. RNN-based models have emerged as the dominant approach for spatiotemporal prediction. As a result, lots of novel RNNs are proposed. For instance, PredRNN [3] introduced a spatiotemporal memory flow to jointly capture spatial and temporal dependencies. MIM [6] employed differential operations to separately model non-stationary and stationary features. MAU [14] incorporated attention mechanisms to enhance information flow by adaptively weighting historical spatiotemporal states. In contrast, CNN models, originally designed for spatial feature extraction, struggle to capture temporal dynamics. However, recent advancements, such as SimVP [9], introduced a translator to learn temporal evolution, and the encoder and decoder are used to learn spatial correlations. Building on this, TAU [10] integrated dynamic inter-frame attention and static intra-frame attention, achieving state-of-the-art performance with a simpler architecture compared to traditional RNN-based approaches.

The Transformer, such as TSViT [12] and PredFormer [13] which possess independent temporal attention layers and spatial attention layers, which model the temporal information and spatial information respectively. This design approach is highly efficient when the video memory is limited.

B. Unified Spatiotemporal Modeling

Transformers with local and global attention have been applied less frequently in spatiotemporal prediction problems. However, compared with the spatiotemporal separation-based Transformer methods, they allow a more flexible token partitioning across the entire spatiotemporal domain. Video Swin Transformer [15] calculates local attention using 3D windows and achieves global attention calculation through shifting 3D windows. Earthformer [4] designs cuboid based attention with different strategies to compute local attention and then uses a global vector to connect local cuboids, enabling global information exchange. Different from the above methods, we separate and fuse tokens to implement local and global attention. Experiments demonstrate that this design can decode and encode spatiotemporal latent variable information, and it is highly efficient for spatiotemporal prediction.

III. Method

Similar to previous works [2], [4], [10], the spatiotemporal sequence forecasting problem is formally defined as follows. Given a spatiotemporal sequence $[\mathcal{X}_i]_{i=1}^{T}, \mathcal{X}_i \in \mathbb{R}^{H\times W\times C}$, we aim to predict the subsequent T' frames $[\mathcal{Y}_{T+i}]_{i=1}^{T'}, \mathcal{Y}_{T+i} \in \mathbb{R}^{H\times W\times C}$, where H , W denote the spatial resolution, and C denotes the channels. In this paper, we train neural networks $\mathcal{F}_\Theta$ parametrized by Θ. Actually, we use stochastic gradient descent to find a set of parameters Θ^* that minimizes the difference between the predicted future frames $\mathcal{F}_\Theta(\mathcal{X}^T)$ and the ground-truth future frames $\mathcal{Y}^{T'}$:

$$\Theta^* = \arg\min_{\Theta} \mathcal{L}(\mathcal{F}_\Theta(\mathcal{X}^T), \mathcal{Y}^{T'}) \tag{1}$$

Where $\mathcal{L}$ is a loss function evaluating differences.

A. Overall Architecture

An overview of the Meld architecture is presented in Fig. 1(a). It first employs a 1×1 convolution to project the number of channels of the input images from C to an arbitrary dimension D. Then, it extracts the spatial features of the images using the Sensation module. In our implementation, D is set as an integer multiple of the number of convolutions in the Sensation module. After the spatial feature extraction, we divide the images into non-overlapping patches. For a patch of size 4×4, its original feature dimension is $4\times4\times D$. A linear embedding layer is applied to this original feature dimension to project it to an arbitrary dimension (denoted as D_l).

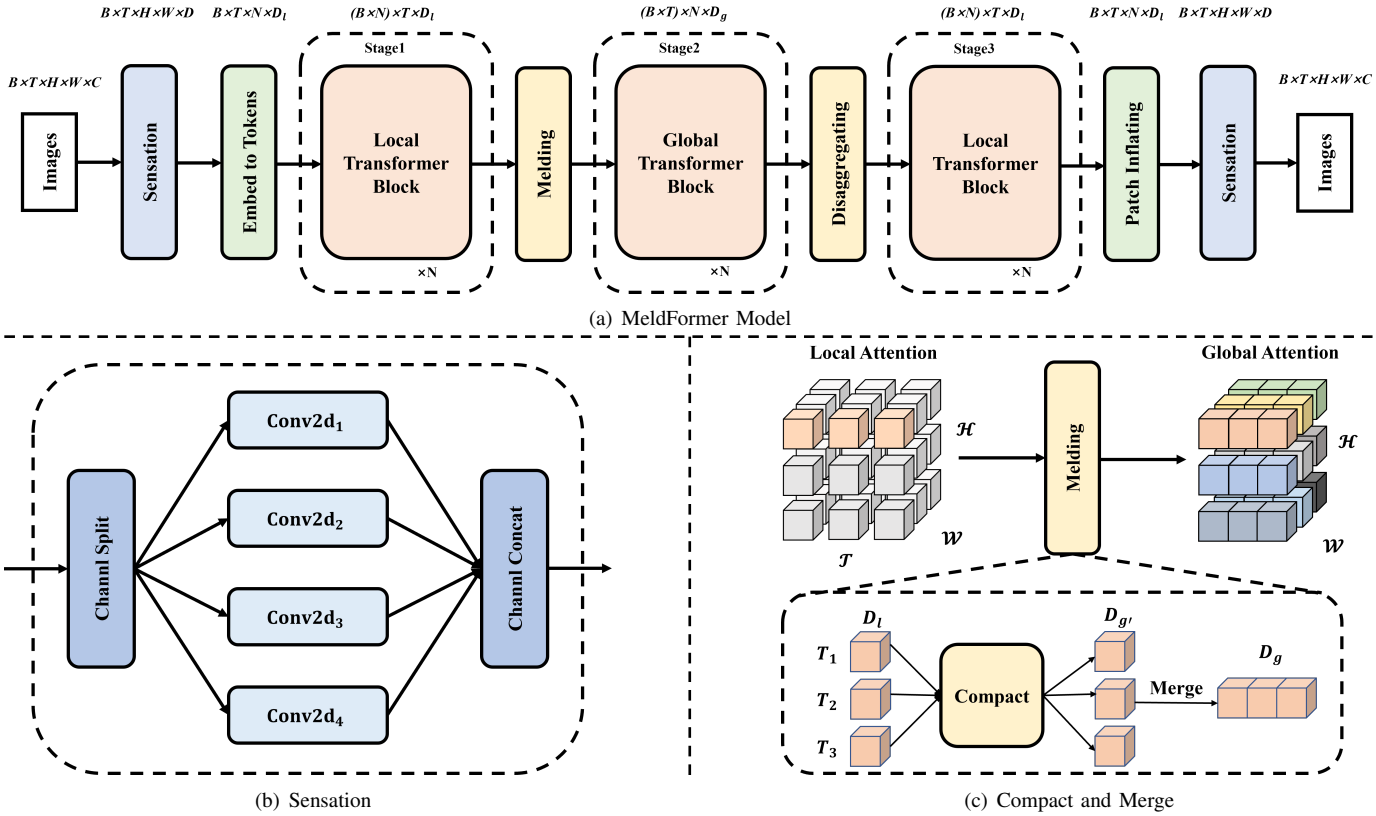
We input the patch tokens at the same position in the images but from different time steps into the Local Transformer blocks for self - attention calculation. This is referred to as the "stage 1". To perform global attention calculation, we compress the dimensions of the tokens and fuse multiple tokens to reduce the number of tokens. Subsequently, the new tokens are input into the Global Transformer blocks for self - attention calculation, which is called the "stage 2".

In the "stage 3", we separate the fused new tokens back to the original number of tokens and restore the token dimensions. Then, we input them into the Local Transformer blocks. We use the Patch Inflating module to project the token dimensions output from the "third stage" back to the original $2D$ patches. Then, we reconstruct the spatial features of the image through the Sensation module. Finally, a 1×1 convolution is used to restore the number of channels of the images to C, obtaining the final predicted frames.

B. Sensation

Since the patches do not overlap with each other during the patch partitioning process, this will lead to an uneven transition of the pixels at the patch edges when reconstructing the image space. Inspired by the works [9], we use the Sensation module for the extraction and reconstruction of the spatial features of images. In this way, not only can the transition of the pixels at the patch edges be smoother and better details be restored during image reconstruction, but also the hidden dimensions of the patch tokens can possess richer spatial information during feature extraction.

As shown in Fig. 1(b), our Sensation module evenly divides the number of image channels according to the number of convolutional blocks, and then inputs them into convolutions with different kernel sizes to extract spatial information. A larger convolutional kernel has a larger receptive field and can better extract global information. Although a smaller convolutional kernel has a smaller receptive field, it can extract more detailed local information. Correspondingly, when reconstructing the spatial features of an image, a large convolutional kernel can restore the overall spatial information, while a small convolutional kernel can better generate local details. Finally, the output features of the convolutional layers with different kernel sizes are concatenated to ensure that the shape



(a) MeldFormer Model

(b) Sensation

(c) Compact and Merge

Fig. 1. (a) The overview architecture of MeldFormer; (b) The specific structure of the Sensation module includes four convolution kernels of different sizes; (c) Compacting and merging tokens: cubes of the same color represent tokens located at the same spatial position but different time steps. D_l denotes the original token dimension, $D_{g'}$ represents the compressed token dimension, and D_g signifies the fused token dimension.

of the image remains unchanged. The Sensation module can be described as:

$$\{\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4\} = \text{Split}(x_{\text{in}}) \tag{2}$$

$$x_{\text{out}} = \text{Concat}\big(\text{Conv2d}_1(\hat{x}_1), \text{Conv2d}_2(\hat{x}_2), \text{Conv2d}_3(\hat{x}_3), \text{Conv2d}_4(\hat{x}_4)\big) \tag{3}$$

C. Melding and Disaggregating Tokens

Performing global attention calculation on the temporal data of images using Transformer is a huge challenge. For the input data tensor with the shape of (T, H, W), the complexity of its self-attention calculation is $O(T^2H^2W^2)$, which is computationally intractable.

Most of the existing solutions adopt spatiotemporal separation modeling [12], [13], which explicitly separates time and space to reduce the computational complexity. Relatively few models address this issue by using a combination of local and global attention, with the design of global attention being a major difficulty. For the two methods, the former requires the temporal layers and the spatial layers to be stacked on each other, while the latter requires a global vector [4] or local displacement [15], and both methods can only achieve the calculation of global attention indirectly.

In order to achieve a more direct calculation of global attention, we compress and fuse the tokens, and then restore the tokens through inflation and decomposition. These processes correspond to the four modules we designed: Local Transformer block, Global Transformer block, Melding, and Disaggregating.

1) Local attention: We first group patch tokens located at the same spatial position across different time steps. Within each group, self-attention is computed independently. Specifically, given an input video sequence $\mathcal{X}_{input} \in \mathbb{R}^{T\times C\times H\times W}$ with T frames, each frame is partitioned into $N = \frac{H}{n} \times \frac{W}{n}$ non-overlapping patches of size $n \times n$. These patches are linearly projected into tokens $[T_i]_{i=1}^{N}$, where each $T_i \in \mathbb{R}^{n\times n\times C}$. By grouping tokens from the same spatial position across all T frames, we form N spatio-temporal token groups $[\mathcal{G}_i]_{i=1}^{N}$, with each group $\mathcal{G}_i \in \mathbb{R}^{T\times n\times n\times C}$ representing the temporal evolution of a specific spatial region. Self-attention is then applied within each group in parallel.

$$\mathcal{G}_i = Grouping_T(\{T_i\}_{t=1}^{T}) \in \mathbb{R}^{T\times n\times n\times C} \tag{4}$$

$$\{\mathcal{G}_i^{local}\}_{i=1}^{N} = LocalAttention(\{\mathcal{G}_i\}_{i=1}^{N}) \tag{5}$$

2) Merge and separate: After local attention computation, we first perform dimensionality compression on the intra-group tokens before fusing them into a single token. Specifically, for the T tokens within a group, a 2D projection matrix is applied to map the original hidden dimension $\text{dim}_{\text{local}}$

to a compressed dimension dim_{new}. The compressed tokens are subsequently concatenated along the hidden dimension in temporal order, yielding a new token $\text{Token}_{\text{global}}$ with a hidden dimension of $T \times \text{dim}_{\text{new}}$. This compression-fusion operation not only reduces the number of intra-group tokens from T to 1 but also transforms the feature dimension from $\text{dim}_{\text{local}}$ to $\text{dim}_{\text{global}}$. Through the merge operation, the new token $\text{Token}_{\text{global}}$ encapsulates the temporal evolution information of spatial features at the same location.

Following the global self-attention computation on $\text{Token}_{\text{global}}$, a decomposition into the original token structure is performed as the inverse process of the compression-fusion operation. Specifically, $\text{Token}_{\text{global}}$ is first evenly partitioned along its hidden dimension into T tokens, each with a hidden dimension of $\frac{\text{dim}_{\text{global}}}{T}$. A 2D projection matrix is then applied to map these dimensions back to the original $\text{dim}_{\text{local}}$, restoring the tokens to their initial shape.

$$\text{Token}_{\text{global}} = Merge(\{\mathcal{G}_i^{local}\}_{i=1}^{N}, \text{dim}_{\text{local}}, \text{dim}_{\text{global}}) \quad (6)$$

$$\{\mathcal{G}_i^{local}\}_{i=1}^{N} = Separate(\text{Token}_{\text{global}}, \text{dim}_{\text{global}}, \text{dim}_{\text{local}}) \quad (7)$$

3) Global attention: After the compression-fusion process, we derive N new tokens termed $\text{Token}_{\text{global}}$, which are subsequently fed into the Global Transformer Block for self-attention computation to enable global information interaction. The hidden dimensions of $\text{Token}_{\text{global}}$ encode the temporal evolution of spatial features at fixed locations, allowing direct cross-correlation of spatio-temporal information across different $\text{Token}_{\text{global}}$ instances during self-attention operations.

In contrast to global attention mechanisms implemented via sliding windows or global vector intermediaries, $\text{Token}_{\text{global}}$ enables direct global information flow without relying on indirect propagation pathways. Additionally, the token count is reduced by a factor of T, drastically achieving more efficient global attention computation.

$$\begin{aligned} X' &= \text{GlobalAttention}(X), \\ X &= \{\text{Token}_{\text{global}}^{1}, \ldots, \text{Token}_{\text{global}}^{N}\} \end{aligned} \quad (8)$$

IV. Experiments

We evaluated our model on commonly used synthetic datasets and real-world datasets, covering a wide range of scenarios, which demonstrates the generalization ability of the model. These scenarios include synthetic handwritten digit trajectories [1], weather forecasting [16], real-world traffic flow [17], and human motion capture [7].

TABLE I
EXPERIMENTAL SETUP. T REPRESENTS THE NUMBER OF FRAMES OF THE INPUT REAL IMAGES, AND T' REPRESENTS THE NUMBER OF FRAMES OF THE PREDICTED OUTPUT IMAGES.

Dataset	C	H	W	T	T'
Moving MNIST	1	64	64	10	10
WeatherBench	1	32	64	12	12
TaxiBJ	2	32	32	4	4
KTH	1	128	128	10	20/40

1) Implementations details: We train our model on an NVIDIA RTX 4090 GPU with 24GB memory. The model employs the L2 loss function and is optimized using the AdamW optimizer. The learning rate is selected from the set $\{1\text{e}-3, 5\text{e}-4\}$ to achieve the best performance, and the OneCycle scheduler is used. Regularization techniques such as Dropout and attention drop are applied to prevent the model from overfitting. Table I lists the more detailed parameters for each dataset.

2) Evaluation metrics: We use four standard metrics to evaluate our model, including the Mean Squared Error (MSE), Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM). We average all results over the predicted frames. Lower MAE/MSE and higher SSIM/PSNR values indicate better prediction performance.

A. Synthetic Datasets

Moving MNIST. The qualitative visualization of prediction results is depicted in Fig. 2(b), while the quantitative outcomes are tabulated in Table II. Our method achieves state-of-the-art performance on the Moving MNIST benchmark, attaining an MSE of 14.1, MAE of 51.2, and SSIM of 0.968. Compared to the TAU [10], our model reduces prediction errors by 28.8% in MSE and 15.1% in MAE. MeldFormer not only successfully predicts the motion trajectories of digits but also preserves their distinct shapes, demonstrating visually indistinguishable quality compared to real sequences.

B. Weather Forecasting

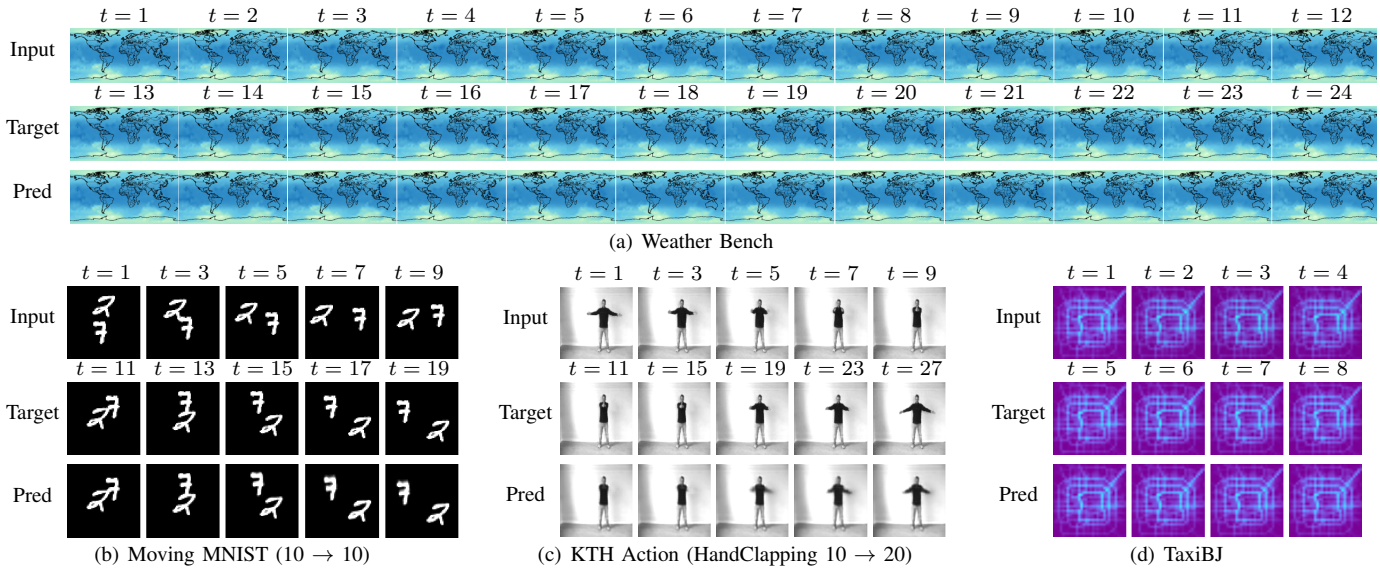
WeatherBench. We show the qualitative visualization in Fig. 2(a) and report the quantitative results in Table II. It is evident that our proposed method achieves state-of-the-art performance on this generalized dataset. Specifically, compared to the leading non-recursive model WaST, our model reduces the MSE from 1.098 to 1.047. These experimental results validate the effectiveness of our model in the domain of climate prediction.

C. Traffic Flow

TaxiBJ. Regarding the prediction results, Fig. 2(d) intuitively presents the qualitative visualization content, and Table II details the quantitative outcomes. Our model surpasses all prior approaches across all evaluated metrics, showcasing robust and dependable predictive capabilities. When compared to the SwinLSTM [26] (an LSTM model integrated with Transformer) and the CNN-based WaST model, MeldFormer achieves notable improvements: MSE is reduced from 0.303 and 0.308 to 0.282, respectively. This demonstrates the superior effectiveness of our model in traffic flow prediction tasks.

D. Human Motion

KTH. Predicting longer frames generally introduces greater prediction difficulty. Compared with SimVP, a CNN-based model, our model improves the PSNR from 33.72 to 34.28, and increases the SSIM from 0.905 and 0.886 to 0.913 and



(a) Weather Bench

(b) Moving MNIST (10 → 10)

(c) KTH Action (HandClapping 10 → 20)

(d) TaxiBJ

Fig. 2. Qualitative results of MeldFormer on four datasets. The first line is the input, the second line is the ground truth, and the third line is the prediction of MeldFormer. Time steps are indicated above each frame. (a) Weather Bench, (b) Moving MNIST, (c) KTH Action, and (d) TaxiBJ.

TABLE II
COMPREHENSIVE QUANTITATIVE COMPARISONS ACROSS DATASETS.

Moving MNIST (10→10)				WeatherBench (12→12)			TaxiBJ (4→4)			Method	KTH (10→20)		KTH (10→40)	
Method	MSE ↓	MAE ↓	SSIM ↑	MSE ↓	MAE ↓	RMSE ↓	MSE ↓	MAE ↓	SSIM ↑		SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑
ConvLSTM [2]	103.3	182.9	0.707	1.521	0.7949	1.233	0.485	17.7	0.978	ConvLSTM [2]	0.712	23.58	0.639	22.85
PredRNN [3]	56.8	126.1	0.867	1.331	0.7246	1.154	0.464	17.1	0.971	SAVP [18]	0.746	25.38	0.701	23.97
PredRNN++ [8]	46.5	106.8	0.898	1.634	0.7883	1.278	0.448	16.9	0.977	DFNDFN [19]	0.794	27.26	0.652	23.01
MIM [6]	44.2	101.1	0.910	1.784	0.8716	1.336	0.429	16.6	0.971	FRNN [20]	0.771	26.12	0.678	23.77
E3D-LSTM [21]	41.3	87.2	0.910	1.592	0.8059	1.233	0.432	16.9	0.979	SV2Pv [22]	0.838	27.79	0.789	26.12
MAU [14]	27.6	86.5	0.937	1.349	0.7391	1.162	-	-	-	PredRNN [3]	0.839	27.55	0.703	24.16
SimVP [9]	23.8	68.9	0.948	1.238	0.7037	1.113	0.414	16.2	0.982	VarNet [23]	0.843	28.48	0.739	25.37
TAU [10]	19.8	60.3	0.957	1.224	0.6810	1.106	0.344	15.6	0.983	SAVP-VAE [24]	0.852	27.77	0.811	26.18
OpenSTL_ViT [25]	19.0	60.8	0.955	1.146	0.6712	1.070	0.317	15.2	0.984	PredRNN++ [8]	0.865	28.47	0.741	25.21
OpenSTL_Swin [25]	18.3	59.0	0.960	1.143	0.6735	1.069	0.313	15.1	0.985	E3D-LSTM [21]	0.879	29.31	0.810	27.24
SwinLSTM [26]	17.7	-	0.962	-	-	-	0.303	15.0	0.984	STMFANet [27]	0.893	29.85	0.851	27.56
WaST [11]	-	-	-	1.098	0.6338	1.044	0.308	14.9	0.984	SimVP [9]	0.905	33.72	0.886	**32.93**
Ours	**14.1**	**51.2**	**0.968**	**1.047**	**0.6335**	**1.023**	**0.282**	**14.5**	**0.986**		**0.913**	**34.28**	**0.895**	32.51

0.895 when predicting 20 and 40 frames based on 10 input frames. The detailed quantitative results are shown in Table II, and visualizations are shown in Fig. 2(c).

E. Ablation Study

In this section, we conduct an ablation study on Moving MNIST trained for 100 epochs to analyze the impact of different modules, and dimensionality compression ratio on model performance.

1) Impact of Different Modules: The experimental results in Table III demonstrate that the GTB enables the flow of global information, indicating that the fused tokens have effectively encoded global spatiotemporal information. Furthermore, the Sensation module captures richer spatial information of images, while the LTB can also effectively decompose the fused tokens.

2) Dimensionality Compression Ratio: Modules ablation experiments demonstrate that the fused tokens can model global spatiotemporal information. Since the dimension of fused tokens is formed by the compressed concatenation of the dimensions of original tokens, we investigate the impact of different compression ratios on model performance. Table IV presents the experimental results under various compression ratios. It can be observed that an excessively high compression

TABLE III
MODULES ABLATION STUDY ON MOVING MNIST DATASET (100 EPOCHS, SHORT TRAINING). GTB REPRESENTS THE GLOBAL TRANSFORMER BLOCK, AND LTB SIGNIFIES THE LOCAL TRANSFORMER BLOCK.

Method	MSE ↓	MAE ↓	SSIM ↑
Ours w/o GTB	103.3	265.0	0.580
Ours w/o Sensation	41.8	106.5	0.903
Ours w/o LTB	33.9	100.6	0.918
Ours	**31.0**	**93.9**	**0.924**

ratio leads to information loss and a relatively high MSE. In contrast, an overly low compression ratio results in a sharp increase in the number of parameters and computational load, without a corresponding reduction in MSE.

TABLE IV
MODEL COMPRESSION PERFORMANCE COMPARISON

Ratio	FLOPs (G)	Parameters	MSE
2:1	45.350	123M	31.58
4:1	**22.104**	**34.82M**	**31.00**
8:1	16.143	12.66M	35.62
16:1	14.578	7.11M	39.90

V. CONCLUSIONS

This paper proposes MeldFormer, a Transformer-based architecture for spatio-temporal prediction. Its core is a token compression-fusion mechanism: grouping co-located tokens across time steps, compressing their dimensions, and fusing them to split attention into intra-group local self-attention (for modeling temporal dynamics of specific spatial regions) and inter-fused-token global self-attention (for capturing cross-spatial correlations). This design shortens the sequence length, thereby reducing computational complexity, while preserving critical spatiotemporal information. Combined with a multi-scale Sensation module for enhanced spatial feature extraction, MeldFormer achieves state-of-the-art performance on Moving MNIST, TaxiBJ, KTH, and WeatherBench across metrics like MSE and SSIM. Ablation studies validate its component effectiveness, demonstrating a balanced solution for spatio-temporal prediction.

REFERENCES

[1] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov, "Unsupervised learning of video representations using lstms," in *International conference on machine learning*. PMLR, 2015, pp. 843–852.

[2] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

[3] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu, "Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms," *Advances in neural information processing systems*, vol. 30, 2017.

[4] Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung, "Earthformer: Exploring space-time transformers for earth system forecasting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25390–25403, 2022.

[5] Shen Fang, Qi Zhang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan, "Gstnet: Global spatial-temporal network for traffic flow prediction.," in *IJCAI*, 2019, pp. 2286–2293.

[6] Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S Yu, "Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9154–9162.

[7] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* IEEE, 2004, vol. 3, pp. 32–36.

[8] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and Philip 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.

[9] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li, "Simvp: Simpler yet better video prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3170–3180.

[10] Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan 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. 18770–18782.

[11] Xuesong Nie, Yunfeng Yan, Siyuan Li, Cheng Tan, Xi Chen, Haoyuan Jin, Zhihang Zhu, Stan Z Li, and Donglian Qi, "Wavelet-driven spatiotemporal predictive learning: bridging frequency and time variations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 4334–4342.

[12] Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou, "Vits for sits: Vision transformers for satellite image time series," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10418–10428.

[13] Yujin Tang, Lu Qi, Fei Xie, Xiangtai Li, Chao Ma, and Ming-Hsuan Yang, "Predformer: Transformers are effective spatial-temporal predictive learners," *arXiv preprint arXiv:2410.04733*, 2024.

[14] Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Yan Ye, Xiang Xinguang, and Wen Gao, "Mau: A motion-aware unit for video prediction and beyond," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26950–26962, 2021.

[15] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[16] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey, "Weatherbench: a benchmark data set for data-driven weather forecasting," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 11, pp. e2020MS002203, 2020.

[17] Junbo Zhang, Yu Zheng, and Dekang Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proceedings of the AAAI conference on artificial intelligence*, 2017, vol. 31.

[18] Hang Gao, Huazhe Xu, Qi-Zhi Cai, Ruth Wang, Fisher Yu, and Trevor Darrell, "Disentangling propagation and generation for video prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9006–9015.

[19] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool, "Dynamic filter networks," *Advances in neural information processing systems*, vol. 29, 2016.

[20] Marc Oliu, Javier Selva, and Sergio Escalera, "Folded recurrent neural networks for future video prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 716–731.

[21] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei, "Eidetic 3d lstm: A model for video prediction and beyond," in *International conference on learning representations*, 2018.

[22] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine, "Stochastic variational video prediction," *arXiv preprint arXiv:1710.11252*, 2017.

[23] Beibei Jin, Yu Hu, Yiming Zeng, Qiankun Tang, Shice Liu, and Jing Ye, "Varnet: Exploring variations for unsupervised video prediction," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 5801–5806.

[24] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine, "Stochastic adversarial video prediction," *arXiv preprint arXiv:1804.01523*, 2018.

[25] Cheng Tan, Siyuan Li, Zhangyang Gao, Wenfei Guan, Zedong Wang, Zicheng Liu, Lirong Wu, and Stan Z Li, "Openstl: A comprehensive benchmark of spatio-temporal predictive learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 69819–69831, 2023.

[26] Song Tang, Chuang Li, Pu Zhang, and RongNian Tang, "Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 13470–13479.

[27] Beibei Jin, Yu Hu, Qiankun Tang, Jingyu Niu, Zhiping Shi, Yinhe Han, and Xiaowei Li, "Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4554–4563.