Self-Distilled Spatiotemporal Mamba Model for Efficient Video Prediction

Jiaxin Li[1], Kai Liu[1,2*], Chunhui Liu[3], Yantao Li[2],
[1] National Elite Institute of Engineering, Chongqing University, Chongqing, China
[2] College of Computer Science, Chongqing University, Chongqing, China
[3] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
lijiaxin987@stu.cqu.edu.cn, liukai0807@cqu.edu.cn, chunhui.liu@polyu.edu.hk, yantaoli@cqu.edu.cn

***Abstract*—Recent advances in deep learning and large-scale models technologies have significantly improved the performance of video prediction. However, it is non-trivial to strike a balance between computational cost and predictive accuracy. In this regard, this work introduces MambaVP, a novel method that embodies the principle of lighter steps for better futures, aiming to achieve higher accuracy of video prediction with lower computational overhead. MambaVP models spatiotemporal dependencies by decoupling spatial and temporal modeling using dimension-specific scanning in STMamba modules. To further ensure that intermediate features retain maximal information useful for predicting the future, we adopt an information-theoretic perspective and introduce a self-distillation strategy, where deep-layer outputs guide shallow-layer representations. This approach effectively minimizes the conditional entropy between the prediction target and intermediate features, thereby reducing uncertainty in future frame prediction. Extensive experiments on MMNIST, Human3.6M, and TaxiBJ demonstrate that MambaVP consistently achieves state-of-the-art (SOTA) performance while significantly reducing computational cost, setting a new benchmark for efficient video prediction.**

***Index Terms*—Video Prediction, Mamba, Self-Distillation**

I. Introduction

Video prediction aims to model spatiotemporal correlations in video data to accurately forecast the future evolution of scenes or events [1]. It has shown promising applications in various domains such as human motion prediction [2], climate prediction [3], and traffic flow prediction [4].

Most existing methods model future frames as the conditional distribution $P(X_{t+1:t+T}|X_{1:t})$, assuming short-term future states can be inferred from observed history. The main challenge is efficiently extracting the information most relevant for predicting future states from high-dimensional video data.

ConvLSTM [5] captures spatial information through convolutions, but its limited receptive field restricts performance. SwinLSTM [6] replaces CNNs with more expressive self-attention mechanisms, improving representational capacity at the expense of increased computational cost. SimVP [7] relies solely on convolutional networks for spatiotemporal feature extraction, significantly reducing computational complexity, but the inherent lack of long-range modeling limits its predictive accuracy. Predformer [8] employs self-attention for spatiotemporal modeling, achieving strong predictive performance, yet the quadratic complexity of self-attention restricts scalability to high-resolution video inputs. Although these approaches make progress along individual axes, few succeed in advancing both performance and efficiency. This raises a fundamental question: *Can we take lighter steps for better futures—designing models that are both accurate and efficient in video prediction?*

We observe that Mamba [9], based on state space models (SSMs), has attracted attention for its strong performance across various tasks [10]. Mamba's RNN-style SSMs provide long-range modeling capability, while its selective scan filters out unimportant inputs to focus on key information. Parallel scan and hardware-aware algorithms enable efficient parallel computation, and near-linear computational complexity keeps the module lightweight. With these strengths, Mamba shows great promise for lightweight, high-performance video prediction. Meanwhile, starting from the fundamental assumption of video prediction, we argue that the core of the task lies in effectively extracting the complete and future-relevant information from past observations. In other words, the extracted representation should contain information that reduces the uncertainty of future frames.

Inspired by this, we propose MambaVP, a fully Mamba-based self-distillation method for video prediction that unifies Mamba and self-distillation. It adopts a spatial–temporal decoupled architecture, leveraging cross scan for spatial representation and Mamba's recurrence to model temporal evolution. To ensure that encoded spatiotemporal features capture information relevant for predicting the future, we employ self-distillation, which aligns shallow-layer features with deep-layer predictions. This effectively reduces the conditional entropy of future frames given the shallow representations, thereby lowering prediction uncertainty conditioned on past observations. Experiments on the TaxiBJ, MMNIST, and Human3.6M datasets demonstrate that MambaVP achieves SOTA with maximal efficiency.

The main contributions of this paper are summarized below:

- We propose MambaVP, a Mamba-based self-distillation method for video prediction, incorporating STMamba for spatiotemporal modeling and leveraging self-distillation to capture information predictive of future frames.
- We design STMamba, a Mamba-based spatiotemporal model with a spatial–temporal decoupled architecture,

*Corresponding author: Kai Liu (email: liukai0807@cqu.edu.cn).

using cross scans for spatial features and unidirectional temporal scanning.
- We reinterpret self-distillation as directional feature regularization. By minimizing the conditional entropy between intermediate representations and predictive targets, the model is encouraged to extract features most informative for future frames.

II. Related Work

A. Video Prediction

Video prediction aims to infer future frames from past frames. Existing methods fall into two categories: explicit and implicit spatiotemporal modeling. Explicit methods rely on optical flow. DMVFN [11] models multi-scale motion via dynamic flow estimation, while OPT [12] formulates prediction as an optimization task, generating bidirectional flows and minimizing appearance and flow consistency losses. Implicit methods learn spatiotemporal representations directly from raw pixels without explicitly modeling motion. VMRNN [13] first incorporates Mamba into an LSTM-style recurrent architecture to capture global spatial dependencies and temporal dynamics. TAU [14] replaces recurrent units with a fully parallel attention mechanism to overcome sequential inference inefficiencies. Modeling capacity and computational efficiency remain key challenges, motivating the development of more efficient and expressive architectures.

B. Self-Distillation

Self-distillation is an effective variant of knowledge distillation that eliminates the need for a separate teacher by using the same network as both teacher and student [15]. Various strategies leverage this concept. Some use deeper layers as internal teachers to guide shallower ones. For example, NSKD [16] introduces auxiliary classifiers in shallow layers to form student–teacher assistant pairs, performing neighbor-level logit distillation to reduce capacity mismatch and improve performance. Others rely on predictions from previous epochs to supervise subsequent learning. TSD [17] mitigates overfitting and enhances generalization by storing prediction results as soft labels and guiding correctly predicted samples with misclassified soft labels. These strategies illustrate self-distillation's flexibility in boosting performance without external teachers.

III. Method

A. Problem Formulation

Given a historical observation sequence $X_t = \{x_1, \ldots, x_t\}$, where each frame $x_i \in \mathbb{R}^{h \times w \times c}$, the goal is to predict the next m frames $Y = \{y_{t+1}, \ldots, y_{t+m}\}$. To formalize this task, we adopt the conditional probability formulation $P(Y|X_t)$, which models the uncertainty of future observations given the past.

Direct prediction in pixel space is challenging due to redundancy and noise. Therefore, we introduce a hierarchical encoder to progressively map X_t into a latent feature hierarchy $\{f_1, f_2, \ldots, f_N\}$, where each layer captures increasingly abstract spatiotemporal dynamics:

$$f_i = \text{encoder}_i(f_{i-1}), \quad f_0 = X_t, \quad i = 1, \ldots, N. \tag{1}$$

The deepest representation f_N serves as a compact latent summary of the historical context and is decoded back to the pixel space to generate the future prediction:

$$\hat{Y} = \text{decoder}(f_N) \tag{2}$$

During the encoding process, each intermediate representation f_i is expected to discard irrelevant redundancy while retaining predictive information about the future Y. Formally, this can be interpreted as maximizing the mutual information $I(Y; f_i)$ across all layers:

$$\max_{\{\text{encoder}_i\}} \sum_{i=1}^{N} I(Y; f_i) \tag{3}$$

This objective ensures that the hierarchical features $\{f_i\}$ progressively encode the most informative cues for future prediction while compressing task-irrelevant input noise.

B. Architecture Overview

An overview of the MambaVP architecture is show in Fig. 1-(a). Each video frame is divided into non-overlapping patches of size $P \times P$ and projected into a feature space via a linear layer. This embedding transforms the 5D temporal sequence input into high-dimensional patch representations capture local spatiotemporal correlations. Spatiotemporal positional encodings are added to patch embedding. The input data is reshaped from $[B, T, C, H, W]$ to $[B, T, h, w, D]$, where B is the batch size, T the number of time steps, C the number of channels, D the embedding dimension, H, W and h, w denote the original and patch-wise spatial dimensions, respectively.

We first describe the deepest prediction path. The embedded patches feed into cascaded STMamba blocks that model spatiotemporal dependencies via multi-directional spatial and temporal scanning. Within each STMamba block, the inputs reshape to $[B \times T, D, h, w]$ and are processed by SpatialMamba to extract frame-level spatial features. These are reshaped to $[B \times h \times w, T, D]$ and passed to TemporalMamba, which captures temporal evolution at consistent spatial locations. To recover original-resolution frames, a recovery layer inversely projects high-dimensional features back to pixel space and reshapes the output to $[B, T, C, H, W]$.

In addition to the main path, we introduce a self-distillation branch. A shallow prediction head is inserted between STMamba blocks, sharing the same architecture as the deepest head and mapping intermediate features to pixel space. During training, outputs from the deepest head supervise these shallow modules, enabling effective self-distillation.

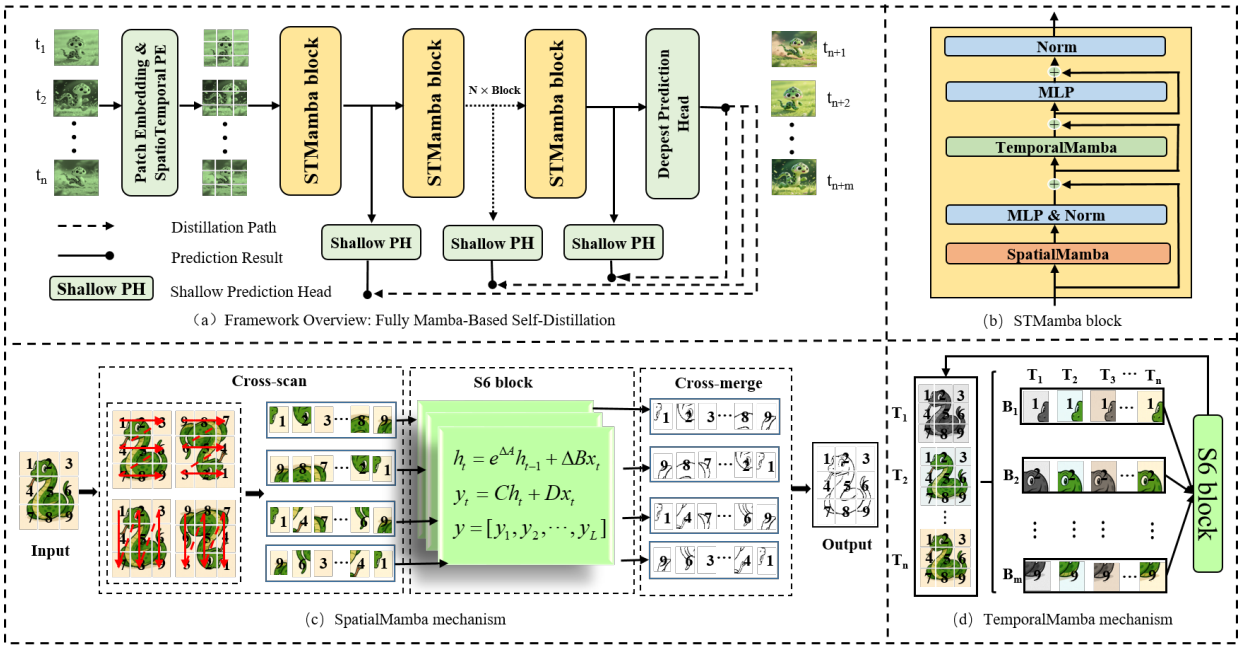

Fig. 1. (a) Overview of the MambaVP model framework. (b) Architecture of the STMamba block. (c) Schematic of the SS2D Multi-directional Scanning Mechanism in SpatialMamba. (d) Schematic of Scanning Mechanism in TemporalMamba.

C. STMamba Block

State space models are a class of continuous systems that map a 1-dimensional function or sequence $x(t) \in \mathbb{R} \mapsto y(t) \in \mathbb{R}$ through an implicit latent state $h(t) \in \mathbb{R}^N$ [9], earliest dates back to the Kalman filter [18]. Continuous-time SSMs can be expressed as linear ordinary differential equations as follows:

$$\begin{aligned} \mathbf{h}'(t) &= \mathbf{A}\mathbf{h}(t) + \mathbf{B}u(t) \\ y(t) &= \mathbf{C}\mathbf{h}(t) + Du(t) \end{aligned} \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$, and $D \in \mathbb{R}^1$ are the weighting parameters.

Since the input tokens in video prediction tasks are discrete, discretization of (4) is required. Specifically, assuming a timescale Δ, a zero-order hold rule can be applied to obtain discrete parameters $\bar{\mathbf{A}} \in \mathbb{R}^{M\times M}$, $\bar{\mathbf{B}} \in \mathbb{R}^{1\times M}$, and $\bar{\mathbf{C}} \in \mathbb{R}^{1\times M}$ according to [19] as shown in Eq. (5):

$$\begin{aligned} \bar{\mathbf{A}} &= \exp(\Delta\mathbf{A}) \\ \bar{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot (\Delta\mathbf{B}) \\ \bar{\mathbf{C}} &= \mathbf{C} \end{aligned} \tag{5}$$

Then the discretized SSMs can be represented as

$$\begin{aligned} h(t) &= \bar{\mathbf{A}}h(t-1) + \bar{\mathbf{B}}x(t) \\ y(t) &= \bar{\mathbf{C}}h(t) \end{aligned} \tag{6}$$

SSMs constitute the core of Mamba, and integrating of selective scan further enhances its predictive capabilities. Notably, the recursive processing inherent to SSMs closely mirrors the evolution of dynamic information in video prediction.

The structure of our proposed Mamba-based spatiotemporal feature extraction module, STMamba, is illustrated in Fig. 1-(b). This module adopts a spatial–temporal decoupled architecture and consists of two main components. First, input data is processed by the spatialMamba block to extract spatial features. These features are then passed through a multilayer perceptron (MLP) layer and a normalization layer to enhance the model's representational capacity and stabilize training. Skip connections are added to facilitate information flow and improve gradient propagation. The process is as follows:

$$x = x + \mathrm{Norm}\Big(\mathrm{MLP}\big(\mathrm{SpatialMamba}(x)\big)\Big) \tag{7}$$

Next, the spatially enriched features are fed into TemporalMamba to model dynamic changes of patches at same spatial location across different time steps. Similarly, MLP, normalization layers, and skip connections are employed in this module to maintain training stability and further enhance model performance. The procedure is as follows:

$$x = \mathrm{Norm}\Big(x + \mathrm{MLP}\big(x + \mathrm{TemporalMamba}(x)\big)\Big) \tag{8}$$

Mamba's unidirectional scanning limits its performance on 2D data. Inspired by VMamba [20], SpatialMamba adopts a cross scan to enhance spatial modeling capability. As illustrated in Fig. 1-(c), our goal is to improve each patch's ability to integrate contextual information from multiple directions, the cross scan divides the input feature map into patch sequences along four traversal paths: top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right. Each sequence is processed in parallel by an S6 block based on SSMs, and the outputs are then reshaped and fused to reconstruct the final spatial feature map.

TemporalMamba adopts a unidirectional scanning strategy along the temporal axis, as temporal information progresses

from past to future. Patches at the same spatial location across time steps exhibit temporal dependencies capture local dynamic changes. As illustrated in Fig. 1-(d), we construct temporal sequences by gathering patches from corresponding spatial locations across time steps. These sequences are processed by S6 block to capture temporal dynamics, and reconstructed into video frames. Inspired by MambaVision Mixer [19], we replace Mamba's causal convolution with standard convolutions, better capture local context, and add a convolutional branch to enhance representational capacity.

D. Self-Distillation

As discussed in III-A, the objective encourages the encoder to focus on extracting features that are informative for predicting the future. To implement this supervision mechanism, we adopt a self-distillation strategy. Specifically, shallow prediction heads are inserted between STMamba blocks to directly project intermediate features into the pixel space. All shallow heads share the same structure as the deepest prediction head, ensuring consistent mapping across layers. During training, the deepest prediction R_d serves as a pseudo target to supervise the shallow predictions R_s, formulated as

$$\mathcal{L}_{\text{sd}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \mathbb{E}\big[\|\mathbf{R}_i - \mathbf{R}_N\|_2^2\big], \tag{9}$$

which encourages intermediate representations to retain information that is predictive of future frames.

Our objective is to generate accurate and visually coherent future predictions. In addition to ensuring that intermediate representations encode information predictive of the future, the final output should also be consistent with the ground-truth future frames. Therefore, the overall training objective combines the self-distillation loss with a reconstruction loss that aligns the predicted frames with the real future frames:

$$\text{Loss} = \mathcal{L}_{\text{main}}(R_d, \text{label}) + \lambda\, \mathcal{L}_{\text{sd}}(R_s, R_d) \tag{10}$$

where $\mathcal{L}_{\text{main}}$ measures the error between the deepest prediction and ground truth, $\mathcal{L}_{\text{sd}}$ is the distillation loss between shallow and deep predictions, both implemented as mean squared error (MSE), and λ balances their contributions.

IV. Experiments

A. Implementation Details

We evaluate our method on three representative benchmark datasets in the video prediction domain: Moving MNIST [24], Human3.6M [25], and TaxiBJ [26]. And we adopt the OpenSTL [27] spatiotemporal prediction framework to carry out the study. We choose the AdamW optimizer to optimize the training of the model with the embedding dimension of 256 for all experiments. The number of training epochs is 200 for Moving MNIST, 50 for Human3.6M, and 200 for TaxiBJ. All experiments are conducted on a single NVIDIA RTX 4090. Following the general guidelines in the field of video prediction, we use the MSE, the mean absolute error (MAE), and the structural similarity index (SSIM) to evaluate the prediction quality. At the beginning of training, predictions from the deepest prediction head are relatively poor. To prevent misleading the shallow modules during the self-distillation stage, we introduce a warmup phase. During this phase, the distillation loss L_{sd} is computed by directly using the ground truth to guide the shallow modules' learning. After the warmup phase, the outputs from the deepest prediction head are employed for self-distillation.

B. Main Results

Quantitative comparisons with previous SOTA methods on three benchmark datasets (Moving MNIST for video prediction, Human3.6M for human motion forecasting, and TaxiBJ for urban flow prediction) are systematically presented in Tab. I and Tab. II, respectively. Experimental results show that the MambaVP achieves better predictive performance while offering clear efficiency advantages, particularly on high-resolution datasets such as Human3.6M. Compared to CNN- and Transformer-based methods, MambaVP reduces computational cost by 6% to 69.0%, and further lowers FLOPs by around 82.4% to 91.4% compared to Recurrent methods.

Building on these efficiency gains, MambaVP also achieves comprehensive superiority over previous state-of-the-art methods across all evaluation metrics, including MSE, MAE, and SSIM, on both the MMNIST and Human3.6M datasets. On the lower-resolution TaxiBJ dataset, the smaller MambaVP-S model excels in MAE and SSIM, whereas the larger MambaVP-L model performs best on MSE. This can be attributed to Mamba's strength in long-range modeling, which is not fully leveraged in low-resolution prediction tasks that require greater focus on local details. Nonetheless, MambaVP achieves competitive overall performance on TaxiBJ compared to existing methods.

In Fig. 2, We present qualitative results of MambaVP on Moving MNIST. The first row shows the input frames, the second row presents the ground truth, the third row displays the predicted results, and the fourth row visualizes the prediction errors. As shown in the figure, MambaVP accurately predicts the motion trajectories of double digits. Even as prediction time increases, the outputs remain sharp without blurring. The error visualization reveals that most errors occur near edges, while overall shapes remain well preserved.

C. Ablation Study

This section conducts ablation studies on the Moving MNIST dataset, focusing on two factors influencing model performance. First, we investigate the impact of different scan strategies used in SpatialMamba module. Second, we evaluate the effectiveness of introducing the self-distillation module.

Scan Strategy We replace the SpatialMamba module's scan mechanism with four variants: Efficient Scan [10], Bidirectional Scan, Local-Global Scan, and Cross Scan. Efficient Scan reduces sequence length via strided patch sampling, recombined into a single frame. Bidirectional Scan unfolds the image in row-major order and generates a reversed sequence. Local-Global Scan uses a hierarchical approach, extracting

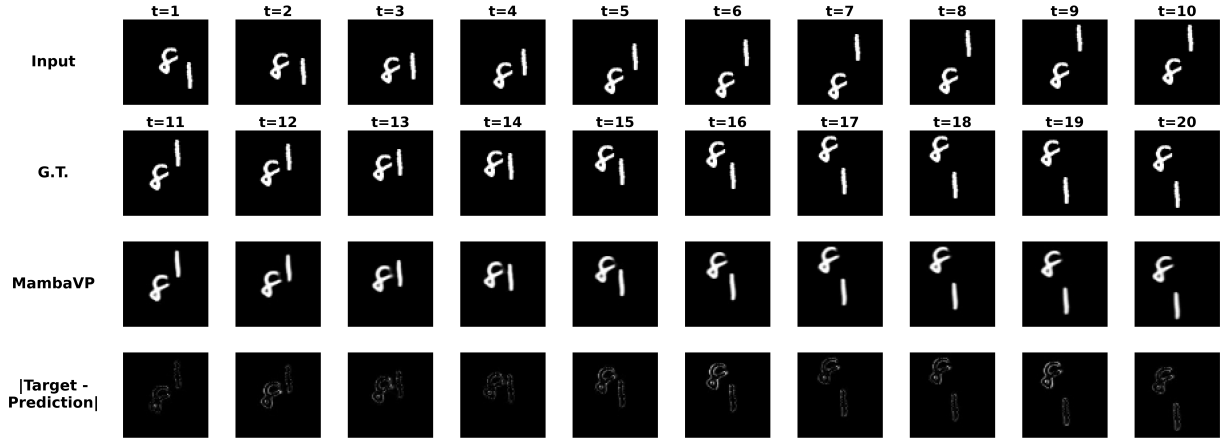

Fig. 2. Qualitative evaluation of the MambaVP model on MMNIST. The first row shows the model input, the second row shows the ground truth, and the third row displays the MambaVP output.

TABLE I

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON MOVING MNIST (200 EPOCHS) AND HUMAN3.6M (50 EPOCHS) DATASETS. THE ARROWS INDICATE THE PREFERRED DIRECTION FOR EACH METRIC: LOWER VALUES (↓) OR HIGHER VALUES (↑) ARE BETTER. DATA MARKED WITH AN ASTERISK * ARE FROM TAU [14] (CVPR 2023). MAMBAVP×10 DENOTES A VARIANT TRAINED FOR 2000 EPOCHS. DETAILS OF ITS PERFORMANCE COMPARED TO THE BASELINE ARE INCLUDED IN THE SUPPLEMENTARY MATERIAL.

Method	Venue	**Moving MNIST**				**Human3.6M**			
		FLOPs $(G)\downarrow$	MSE↓	MAE↓	SSIM↑	FLOPs $(G)\downarrow$	MSE $/10\downarrow$	MAE $/100\downarrow$	SSIM↑
ConvLSTM [5]	NIPS'2015	56.8	103.3	182.9	0.707	347.0	12.55	15.67	0.981
E3D-LSTM [21]	ICLR'2018	298.9	41.3	86.4	0.910	542.0	14.33	14.43	0.980
SimVP [7]	CVPR'2022	19.4	32.2	89.1	0.927	197.0	31.60	15.12	0.904
PredRNNv2 [22]	TPAMI'2022	116.6	27.7	82.2	0.937	708.0	11.49	14.85	0.983
SwinLSTM [6]	ICCV'2023	69.9	27.4	78.7	0.938	-	11.90	33.20	0.913
PredFormer [8]	arXiv'2024	16.5	27.8	80.5	0.937	65.0	11.27	13.86	**0.984**
SimVPv2 [23]	TMM'25	16.5	26.6	77.3	0.940	182.0*	11.33*	13.91*	**0.984***
MambaVP	Ours	**15.8**	**25.9**	**75.0**	**0.944**	**61.0**	**10.92**	**13.77**	**0.984**
MambaVP×10	Ours	15.8	14.3	48.0	0.970	-	-	-	-

TABLE II

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON TAXIBJ DATASET. MAMBAVP-S AND MAMBAVP-L DENOTE SMALL AND LARGE PARAMETER MODELS, RESPECTIVELY.

Method	FLOPs(G)↓	MSE $\times 100\downarrow$	MAE↓	SSIM↑
ConvLSTM [5]	20.7	48.5	17.7	0.978
E3D-LSTM [21]	98.2	43.2	16.9	0.979
SimVP [7]	3.6	41.4	16.2	0.982
PredRNNv2 [22]	-	45.2	16.8	0.980
IAM4VP [28]	-	37.2	16.4	0.983
PredFormer [8]	**2.2**	27.7	**14.3**	0.986
SimVPv2 [23]	2.6	34.8	15.6	0.984
MambaVP-S	2.5	28.5	**14.3**	**0.987**
MambaVP-L	3.8	**27.2**	14.4	**0.987**

TABLE III

ABLATION STUDY OF SCAN STRATEGIES ON MMNIST.

Scan Strategy	FLOPs (G)↓	MSE↓	MAE↓	SSIM↑
Efficient Scan	**12.9**	27.9	80.9	0.937
Bidirectional Scan	14.5	28.1	82.2	0.937
Local-Global Scan	13.7	29.0	83.8	0.935
Cross Scan	15.8	**26.8**	**79.2**	**0.940**

local features within quadrants before performing a global scan to capture context.

As shown in Tab. III, the Bidirectional and Local-Global Scan strategies improve computational efficiency but significantly degrade prediction accuracy. The Efficient Scan has the lowest cost but still compromises accuracy. In contrast, Cross Scan preserves spatial structure, achieving better performance through multi-directional context integration.

Self-Distillation Ablation To validate the effectiveness of our Mamba-based self-distillation framework, we vary the hyperparameter λ in Eq. 10 to control distillation strength. As shown in Tab. IV, self-distillation consistently improves performance over the $\lambda = 0$ baseline, but excessive strength reverses this gain. These results suggest that impos-

TABLE IV
ABLATION STUDY OF SELF-DISTILLATION STRENGTH ON MMNIST PERFORMANCE.

Distillation Strength λ	MSE↓	MAE↓	SSIM↑
0.0	26.81	79.22	0.940
0.5	26.68	77.83	0.941
0.7	26.22	76.23	0.943
1.1	**25.86**	**75.04**	**0.944**
1.2	25.90	75.05	0.933

ing information-theoretic constraints on intermediate features compresses their representation space, retaining components relevant to the prediction target and reducing feature noise, thereby enhancing accuracy. However, overly strong distillation may cause shallow features to lose diversity or local detail, or propagate deep-layer errors into shallow representations, degrading shallow-layer performance.

V. CONCLUSION

In this work, we proposed MambaVP, a novel fully Mamba-based self-distillation method for video prediction. Our method introduces STMamba, a Mamba-based architecture that decouples spatial and temporal modeling to efficiently capture spatiotemporal dependencies. To ensure that encoded spatiotemporal features capture information relevant for predicting the future, we incorporate a self-distillation mechanism in which deeper-layer representations are used to guide earlier layers, enhancing feature learning and prediction accuracy. Experimental results demonstrate that MambaVP achieves SOTA on multiple benchmark datasets while significantly reducing computational cost, making it well-suited for deployment on resource-constrained devices. In the end, our work takes lighter steps for better futures, steadily advancing toward more efficient and accurate predictive frameworks.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 62472055, the Guangdong Basic and Applied Basic Research Foundation under Grant 2022B1515020015, and Chongqing Key Laboratory of Big Data Intelligence and Privacy Computing.

REFERENCES

[1] S. Oprea, P. Martinez-Gonzalez, A. Garcia-Garcia, J. A. Castro-Vargas, S. Orts-Escolano, J. Garcia-Rodriguez, and A. Argyros, "A review on deep learning techniques for video prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2806–2826, 2020.

[2] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *Computer vision and image understanding*, vol. 171, pp. 118–139, 2018.

[3] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate medium-range global weather forecasting with 3d neural networks," *Nature*, vol. 619, no. 7970, pp. 533–538, 2023.

[4] 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.

[5] S. Xingjian, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, p. 1, 2015.

[6] 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.

[7] 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.

[8] Y. Tang, L. Qi, F. Xie, X. Li, C. Ma, and M.-H. Yang, "Predformer: Transformers are effective spatial-temporal predictive learners," *arXiv preprint arXiv:2410.04733*, 2024.

[9] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[10] X. Pei, T. Huang, and C. Xu, "Efficientvmamba: Atrous selective scan for light weight visual mamba," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 6443–6451.

[11] X. Hu, Z. Huang, A. Huang, J. Xu, and S. Zhou, "A dynamic multi-scale voxel flow network for video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6121–6131.

[12] Y. Wu, Q. Wen, and Q. Chen, "Optimizing video prediction via video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 814–17 823.

[13] 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*, 2024, pp. 5663–5673.

[14] C. Tan, Z. Gao, L. Wu, Y. Xu, J. Xia, S. Li, and S. Z. Li, "Temporal attention unit: Towards efficient spatiotemporal predictive learning," 2023. [Online]. Available: https://arxiv.org/abs/2206.12126

[15] A. M. Mansourian, R. Ahmadi, M. Ghafouri, A. M. Babaei, E. B. Golezani, Z. Y. Ghamchi, V. Ramezanian, A. Taherian, K. Dinashi, A. Miri *et al.*, "A comprehensive survey on knowledge distillation," *arXiv preprint arXiv:2503.12067*, 2025.

[16] P. Liang, W. Zhang, J. Wang, and Y. Guo, "Neighbor self-knowledge distillation," *Information Sciences*, vol. 654, p. 119859, 2024.

[17] M. Liu, Y. Yu, Z. Ji, J. Han, and Z. Zhang, "Tolerant self-distillation for image classification," *Neural Networks*, vol. 174, p. 106215, 2024.

[18] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

[19] A. Hatamizadeh and J. Kautz, "Mambavision: A hybrid mamba-transformer vision backbone," *arXiv preprint arXiv:2407.08083*, 2024.

[20] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," *Advances in neural information processing systems*, vol. 37, pp. 103 031–103 063, 2024.

[21] 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.

[22] 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," 2022. [Online]. Available: https://arxiv.org/abs/2103.09504

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

[24] 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.

[25] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[26] 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.

[27] 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.

[28] M. Seo, H. Lee, D. Kim, and J. Seo, "Implicit stacked autoregressive model for video prediction," 2023. [Online]. Available: https://arxiv.org/abs/2303.07849