Towards Robust Deterministic and Probabilistic Modeling for Predictive Learning

Xuesong Nie[1,2] , **Haoyuan Jin**[1] , **Vijayakumar Bhagavatula**[3] and **Xiaofeng Liu**[2,*]
[1]Zhejiang University
[2]Yale University
[3]Carnegie Mellon University
{xuesongnie, jhyjhy}@zju.edu.cn, kumar@ece.cmu.edu, xiaofeng.liu@yale.edu

Abstract

Predictive modeling of unannotated spatiotemporal data presents inherent challenges, primarily due to the highly entangled visual dynamics in real-world scenes. To tackle these complexities, we introduce a novel insight through Disentangling Deterministic and Probabilistic (DDP) modeling. We note a key observation in spatiotemporal data where low-level details typically remain stable, whereas high-level motion frequently exhibits dynamic variations. The core motivation involves constructing two distinct pathways in the latent space: a deterministic path and a probabilistic path. The probabilistic path begins by defining the motion flow, which explicitly describes complex many-to-many motion patterns between patches, and models its probabilistic distribution using a motion diffuser. The deterministic path incorporates a spectral-aware enhancer to retain and amplify visual details in the frequency domain. These designs ensure visual consistency while also capturing intricate long-term motion dynamics. Extensive experiments demonstrate the superiority of DDP across diverse scenario evaluations.

1 Introduction

Predictive learning, a self-supervised learning method, excels in uncovering latent structures within unannotated spatiotemporal data. This topic models temporal evolution by predicting future frames from given ones, offering extensive applications in autonomous driving [Jin *et al.*, 2024], climate modeling [Lam *et al.*, 2022], traffic flow [Nie *et al.*, 2024d], robotics [Gupta *et al.*, 2022], and popular world simulations [Nie *et al.*, 2024b]. However, spatiotemporal sequences, abundant and readily available in nature, typically exhibit intricate spatial correlations, movement trends, and multi-object interaction in practical scenarios.

Struggling with the inherent complexity and randomness of future events, predictive learning has developed into two main approaches: *recurrent-based* and *recurrent-free*. Recurrent-based methods consist of recurrent unit variants

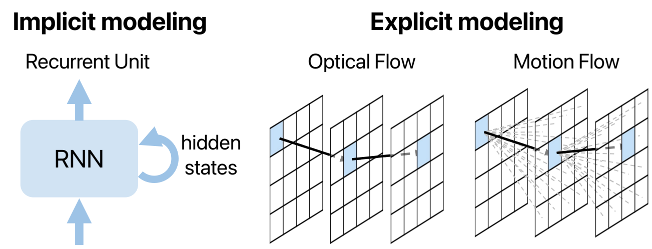

Figure 1: Illustration of two motion modeling approaches. The implicit modeling methods using recurrent units or spatiotemporal translators embed motion information into hidden features. The optical flow describes the one-to-one mapping of pixels between different frames. Our proposed motion flow captures many-to-many motion patterns by computing the similarity of latents across frames.

(*e.g.*, LSTM [Hochreiter and others, 1997], ConvLSTM [Shi *et al.*, 2015], and ST-LSTM [Wang *et al.*, 2017]) and state transition connections across timesteps. These methods embed motion information into hidden states to model temporal dynamics, as shown in Fig. 1. Recurrent-free methods employ parallel spatiotemporal translators instead of recurrent units to model spatiotemporal dependencies. SimVP [Gao *et al.*, 2022], TAU [Tan *et al.*, 2023], and TAT [Nie *et al.*, 2024a] propose elaborate spatiotemporal learning modules for implicit temporal evolution capture. These models are required to adeptly acquire sophisticated motion patterns autonomously. DMVFN [Hu *et al.*, 2023], on the other hand, utilize optical flow to gracefully refine explicit motion depiction, effectively diminishing the occurrence of artifacts. Nonetheless, these methods inevitably exhibit limitations over time, as shown in Fig. 2. DMVFN [Hu *et al.*, 2023] captures motion effectively but often fails to maintain visual consistency, yielding "correct" yet not "ideal" results, *e.g.*, more details of clothes or horse heads are lost. Conversely, TAU [Tan *et al.*, 2023] excels in preserving detailed visual features but struggles with accurate motion representation, *e.g.*, the motion of the arm produces deformation.

The observations above reveal that low-level details typically remain stable, whereas high-level motion exhibits dynamic variations. This seems intuitive, for example, an individual's visual features remain relatively constant in the short term, while their behavior exhibits high stochasticity. The prediction can be improved by designing models based on

[*]Corresponding Author.

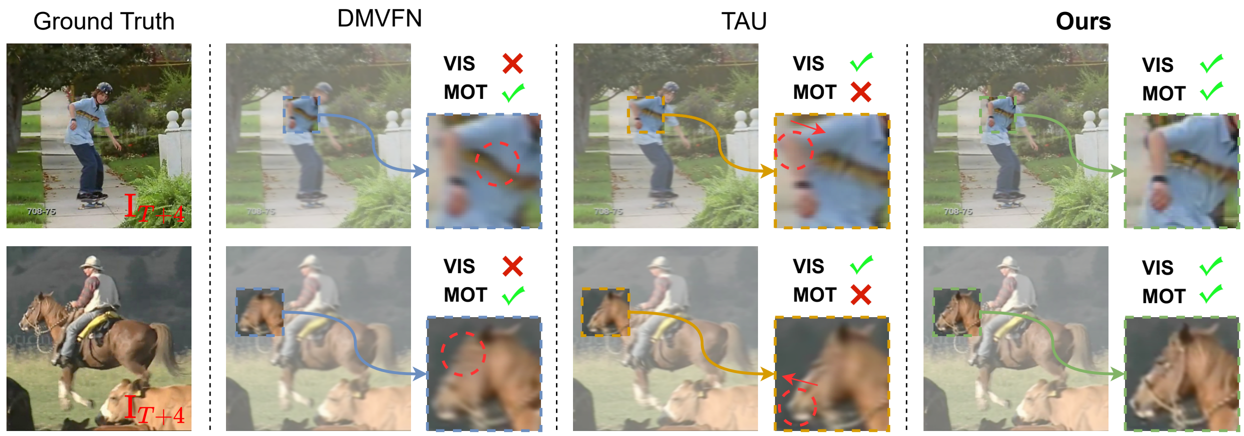

Figure 2: Comparative results of different methods on the UCF Sports datasets at the $(T+4)_{th}$ frame. VIS is short for visual appearance, and MOT for motion dynamics. Our DDP models motion dynamics well while retaining more detailed features.

these data properties, which existing methods have not fully explored. To this end, we address two key questions:

(i) *How to represent and model high-level motion dynamics?* Explicit modeling is essential for understanding complex motion dynamics. The most common methods like optical flow describe a one-to-one pixel mapping with two channels, but this approach often overlooks interactions between more pixels, as shown in Fig. 1. Therefore, we formulate motion flow by computing inter-frame latent similarities, creating a many-to-many mapping with more channels. Although motion flow captures richer motion patterns, its long-term motion distribution modeling is more challenging. To address this, we introduce a motion diffuser that employs a spatiotemporal state space model with the diffusion structure to learn the transition from Gaussian noise to true motion distribution.

(ii) *How to enhance low-level details and multi-scale dynamics?* Most of the leading methods typically extract features in the original latent space. In contrast, we propose a spectral-aware enhancer that models each frame individually in the frequency domain, preserving more visual detail features. To address significant cross-scale motion variations between consecutive frames, we implement motion flow sharing and motion-visual warping strategies. These approaches significantly reduce complexity and enhance long-term prediction capabilities.

In this paper, we present a new perspective through Disentangling Deterministic and Probabilistic (DDP) modeling for robust predictive learning. Extensive experiments demonstrate the effectiveness of DDP for various prediction scenarios. Our key contributions are summarized as follows:

- Introduced DDP, a novel framework that disentangles deterministic visuals from probabilistic motion for spatiotemporal prediction.
- Proposed explicit motion modeling (motion flow) and a spatiotemporal diffusion model to capture its complex, probabilistic nature.
- Developed a spectral-aware enhancer and motion-visual warping techniques to improve low-level detail preservation and handle multi-scale dynamics.
- Extensive experiments show that DDP achieves state-of-the-art performance across various real-world scenes.

2 Related Work

2.1 Predictive Learning Models

Recurrent-based models have historically dominated predictive learning. ConvLSTM [Shi *et al.*, 2015] augments LSTM's spatial learning with convolutional architecture. PredRNN [Wang *et al.*, 2017] introduces spatiotemporal LSTM, capturing spatial and temporal dependencies. PredRNN++[Wang *et al.*, 2018b] addresses gradient vanishing through a gradient highway unit. MCnet [Villegas *et al.*, 2017] decomposes the motion and content modeling with LSTM and CNN. SADM [Bei *et al.*, 2021] fuses the content semantic maps and optical flow motion maps for future frame prediction. E3D-LSTM [Wang *et al.*, 2018c] extends LSTM with 3D convolution. PredRNNv2 [Wang *et al.*, 2022] employs curriculum learning and memory decoupling loss. WaST [Nie *et al.*, 2024d] presents an innovative wavelet-based spatiotemporal framework for modeling spatial frequency and temporal variations. ModeRNN [Yao *et al.*, 2023] uses spatiotemporal slots to extract visual dynamics components, addressing mode collapse.

Recent research has shifted towards recurrent-free models to overcome parallelization limitations. vid2vid [Wang *et al.*, 2018a] decomposes video visuals and motion for frame prediction with spatiotemporal adversarial learning. [Wu *et al.*, 2020] decomposes the background scene and moving objects with instance maps. SimVP [Gao *et al.*, 2022] employs Inception modules with UNet architecture for temporal dependency learning. TAU [Tan *et al.*, 2023] decomposes temporal attention into intra-frame static and inter-frame dynamical components. DMVFN [Hu *et al.*, 2023] proposes a dynamic multi-scale voxel flow network. In contrast, our approach disentangles deterministic visuals and probabilistic motion in latent space to enhance prediction accuracy.

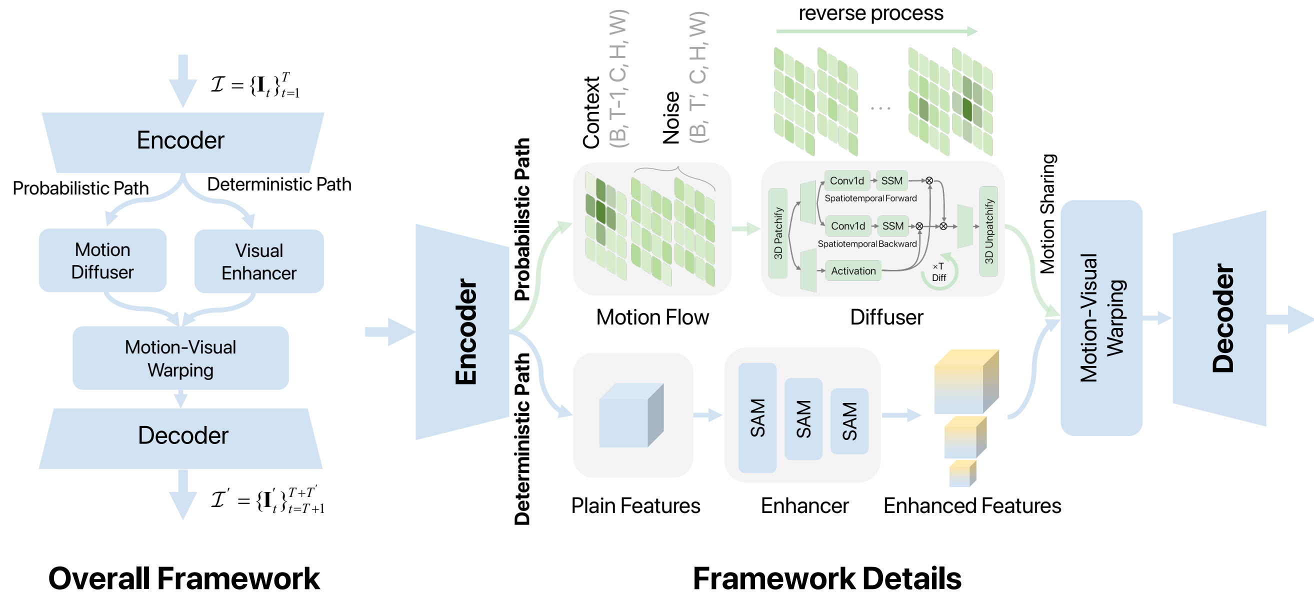

Figure 3: Illustration of the DDP architecture. Our DDP features two pathways in latent space: the deterministic path and the probabilistic path. The probabilistic path models the distribution of well-defined motion flows through a motion diffuser, while the deterministic path preserves and enhances visual details in the frequency domain via a spectral-aware enhancer.

2.2 State Space Models

State Space Models (SSMs) with efficient hardware-aware designs, *e.g.*, Mamba [Gu and Dao, 2023], have recently demonstrated significant potential for long sequence modeling with linear complexity. ViS4mer [Islam and Bertasius, 2022] employs a 1D Structured State Space Sequence (S4) model for long-range temporal dependencies in video classification. S4ND [Nguyen *et al.*, 2022] extends 1D S4 to multi-dimensional data, including 2D images and 3D videos. TranS4mer [Islam *et al.*, 2023] combines self-attention and S4 for movie scene detection, while S5 [Wang *et al.*, 2023] introduces a selective mechanism to S4, enhancing its performance in long-form video understanding. DiffuSSM [Yan *et al.*, 2023] replaces attention mechanisms with a more scalable SSM-based backbone for high-resolution image generation. ViM [Zhu *et al.*, 2024] demonstrates that self-attention is not essential for visual representation learning by constructing a pure SSM-based model. VMamba [Liu *et al.*, 2024] addresses the direction-sensitive issue by introducing a cross-scan module to traverse the spatial domain. Our work explores spatiotemporal diffusion SSMs for probabilistic modeling of motion dynamics.

3 Method

3.1 Framework Overview

The spatiotemporal predictive learning aims to model spatial and temporal dependencies of given past T frames $\mathcal{I} = \{\mathbf{I}_t\}_{t=1}^{T}$ to predict the most reasonable future T' frames $\mathcal{I}' = \{\mathbf{I}'_t\}_{t=T+1}^{T+T'}$, where $\mathbf{I}_t \in \mathcal{R}^{C\times H\times W}$ denotes the t_{th} frame. Our DDP involves three important components: (i) Probabilistic Motion Modeling, (ii) Deterministic Visual Modeling, and (iii) Motion-Visual Warping. We detail them in the following.

3.2 Probabilistic Motion Modeling

We first construct the visual-agnostic motion flow $\{\mathbf{F}_{t\to t+1}\}, t \in \{1, 2, \ldots, T-1\}$ by computing the token similarity across frames. Then the motion diffuser models past motion flow $\mathrm{MDiff}(\mathbf{F}_{t\to t+1})$ to estimate future motion flow $\hat{\mathbf{F}}_{T\to T+t'}$ in an iterative manner.

Motion Flow Formulation. Given the latent feature maps $\mathbf{X}_t \in \mathcal{R}^{C\times N}$, t ranges from 1 to T and $N = H\times W$. As illustrated in Fig. 4(a), we represent the i_{th} feature patch as $\mathbf{X}_t^i$ and compute dot product similarity for two consecutive frames $\{\mathbf{X}_t, \mathbf{X}_{t+1}\}$ to formulate the motion flow:

$$\mathbf{F}_{t\to t+1}^{i,j} = \mathrm{Sim}(\mathbf{X}_t^i, \mathbf{X}_{t+1}^j), \forall i, j \in \{0, \ldots, N-1\}. \quad (1)$$

Unlike optical flow that establishes one-to-one pixel correspondence, motion flow $\mathbf{F}_{t\to t+1}^{i,j}$ captures the many-to-many patch relationship across frames, showing the impact of the i_{th} patch on j_{th} patch in different frames.

Motion Flow Estimation. To effectively capture long-term motion patterns, we explored a new spatiotemporal diffusion state space model, motion diffuser, which is constructed from a sequence of SSMs. They are systems that map a 1D function or sequence $x(t) \in \mathcal{R} \mapsto y(t) \in \mathcal{R}$. It can be expressed as a linear Ordinary Differential Equation (ODE):

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), y(t) = \mathbf{C}h(t), \quad (2)$$

where $\mathbf{A} \in \mathcal{R}^{M\times M}$ and $\mathbf{B}, \mathbf{C} \in \mathcal{R}^{M}$ are its parameters and $h(t) \in \mathcal{R}^M$ denotes the hidden state. The discrete versions [Karras *et al.*, 2022] of ODE include a timescale parameter $\mathbf{\Delta}$ to transform the continuous parameters $\mathbf{A}, \mathbf{B}$ to

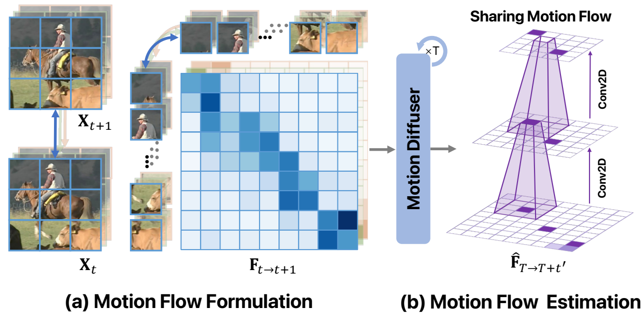

Figure 4: Illustration of motion flow formulation and estimation process. (a) The motion flow is formulated by computing token similarity across frames. (b) The motion diffuser estimates future motion flow in an iterative manner.

discrete parameters $\overline{\mathbf{A}}, \overline{\mathbf{B}}$. One common discrete method is the Zero-Order Hold (ZOH), represented as:

$$\overline{\mathbf{A}} = \exp(\mathbf{\Delta A}), \overline{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I})\mathbf{\Delta B}, \quad (3)$$

$$h_t = \overline{\mathbf{A}} h_{t-1} + \overline{\mathbf{B}} x_t, y_t = \mathbf{C} h_t. \quad (4)$$

Mamba [Gu and Dao, 2023] further extends the discretization process with a selection mechanism. Based on these, we build the motion diffuser using stacked SSM blocks. Given the motion flow of the past T frames $\{\mathbf{F}_{t\to t+1}\} \in \mathcal{R}^{(T-1)\times N\times C}$, where $C = N$. These sequences are input into the motion diffuser to perform the forward and backward spatiotemporal selective scan, as shown in Fig. 3. Specifically, the sequences are aggregated after the forward scan and the reverse scan of the flipped sequences. Finally, the estimated motion flow can be formulated as:

$$\hat{\mathbf{F}}_{T\to T+t'} = \mathrm{MDiff}(\mathbf{F}_{t\to t+1}), \forall t \in \{1, 2, \ldots, T-1\}, \quad (5)$$

where we directly estimate the motion flow from the T_{th} to the $(T+t')_{th}$ frame to reduce error accumulation. Moreover, as shown in Fig. 4(b), the MDiff cross-scale shares the estimated motion flow through convolutional projection to reduce multi-scale iterations. The diffusion process progressively adds noise to the motion flow, as shown in Fig. 5. Let β_t represent the noise variance ratio at time t, and $\alpha_t = 1-\beta_t$. With context motion flow as condition c and denoised motion flow as z, these are concatenated as input. The training loss for the motion diffuser $\epsilon_\phi\left(z^t; t\right)$ is:

$$\mathcal{L}\left(\epsilon_\phi\right) = E_{t,c}\left\|\epsilon_\phi\left(\sqrt{\bar{\alpha}_t}z + \sqrt{1-\bar{\alpha}_t}\epsilon, t, c\right) - \epsilon\right\|^2. \quad (6)$$

3.3 Deterministic Visual Modeling

While recurrent-based models excel at capturing motion patterns, they often struggle to maintain visual consistency. To address this limitation, we introduce the Spectral-Aware Module (SAM) designed to enhance frequency-domain representations (as shown in Fig. 6(a)). Adopting a MetaFormer-like paradigm [Yu *et al.*, 2022; Nie *et al.*, 2024c], SAM comprises: (i) Dilated Reparam Convolution (DRConv) [Ding *et al.*, 2023] for token mixing, which augments a non-dilated large kernel with parallel reparameterizable dilated small kernels, and (ii) Energy-based Frequency Channel Mixing

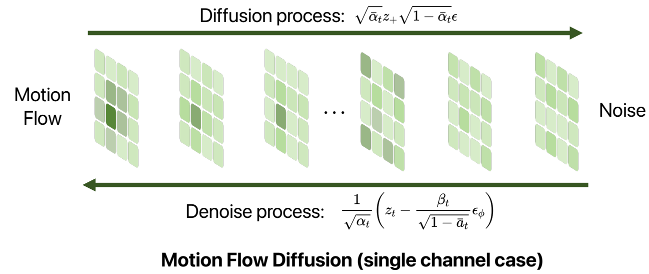

Figure 5: Illustration of the single-channel diffusion process. The vector is reshaped into a spatial format for visualization.

(EFCM) for channel mixing. EFCM computes the mean $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and variance $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \hat{\mu}\right)^2$ of latent features $\mathbf{X}$, where $N = H \times W$. The energy value [Yang *et al.*, 2021] is determined by minimizing:

$$e_{i,j} = \frac{4\left(\hat{\sigma}^2 + \lambda\right)}{(t_{i,j} - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\delta}, \quad (7)$$

where δ represents the hyper-parameter, and $e_{i,j}$ is the energy value of the target token $t_{i,j}$, where i ranges from 0 to $H-1$, and j ranges from 0 to $W-1$. The energy feature $\mathbf{X}_e$ is formed by grouping all $e_{i,j}$ values. Then, we rescale the energy value to limit excessively large energy values:

$$e^*_{i,j} = \mathrm{LeakyReLU}\left(\frac{1}{\sum_{h,w} e_{i,j}}\right). \quad (8)$$

For energy features, we pool it into a global vector $x \in \mathcal{R}^{C\times 1\times 1}$, and then transform it to Fourier space:

$$\mathcal{F}(x)(z) = \frac{1}{\mathrm{C}}\sum_{\mathrm{c}=0}^{\mathrm{C}-1} x(\mathrm{c})\mathrm{e}^{-\mathrm{j}2\pi\frac{\mathrm{c}}{\mathrm{C}}z}, \quad (9)$$

where amplitude component $\mathcal{A}(x)(z)$ and phase component $\mathcal{P}(x)(z)$ of $\mathcal{F}(x)(z)$ represent different information, thus we introduce attention-based operations to enhance $\mathcal{A}(x)(z)$ and $\mathcal{P}(x)(z)$ respectively:

$$\mathcal{A}\left(x\right)\left(z\right)' = \mathcal{A}_{\text{filter}}\left(x\right) \odot \mathcal{A}\left(x\right)\left(z\right), \quad (10)$$

$$\mathcal{P}\left(x\right)\left(z\right)' = \mathcal{P}_{\text{filter}}\left(x\right) \odot \mathcal{P}\left(x\right)\left(z\right), \quad (11)$$

where $\mathcal{A}_{\text{filter}}(\cdot)$ and $\mathcal{P}_{\text{filter}}(\cdot)$ denote 1×1 filters for the amplitude and phase components. The symbol $\odot$ signifies the Hadamard product for attention weighting. Then we convert the Fourier features to their original space via the inverted Fourier transform $\mathcal{F}^{-1}\left(\mathcal{A}\left(x\right)\left(z\right)', \mathcal{P}\left(x\right)\left(z\right)'\right)$. Finally, we context broadcast it to the original input.

3.4 Motion-Visual Warping

Inspired by the warping operation in optical flow, which uses the flow $\mathbf{F}_{t\to t+1}$ to map $\mathbf{I}_t$ to $\mathbf{I}_{t+1}$ in pixel space. Similarly, the motion-visual warping applies motion flow $\hat{\mathbf{F}}_{T\to T+t'}$ on the observed T visual features $\mathbf{X}_t$ to obtain the $(T+t')_{th}$ features $\hat{\mathbf{X}}_{T+t'}$ in latent space:

$$\hat{\mathbf{X}}_{T+t'} = \left(\sum_{t=1}^{T}\mathbf{X}_t \cdot \mathbf{F}_{t\to T}\right) \cdot \hat{\mathbf{F}}_{T\to T+t'}. \quad (12)$$

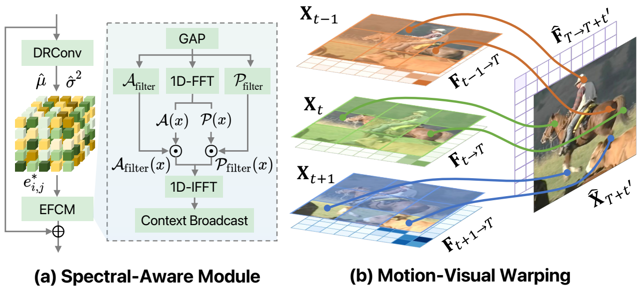

Figure 6: (a) Spectral-aware module architecture involves RDD-Conv for token mixing, and energy-based frequency channel mixing. (b) Motion-visual warping aggregates past features to generate credible future features.

Method	$T+1$		$T+6$	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
ConvLSTM [Shi *et al.*, 2015]	26.43	32.20	17.80	58.78
PredRNN [Wang *et al.*, 2017]	27.17	28.15	19.65	55.34
PredRNN++ [Wang *et al.*, 2018b]	27.26	26.80	19.67	56.79
E3D-LSTM [Wang *et al.*, 2018c]	27.98	25.13	20.33	47.76
MotionRNN [Wu *et al.*, 2021]	27.67	24.23	20.01	49.20
STRPM [Chang *et al.*, 2022]	28.54	20.69	20.59	41.11
SimVP [Gao *et al.*, 2022]	30.64	13.17	21.83	38.74
DMVFN [Hu *et al.*, 2023]	30.05	10.24	22.67	22.50
WaST [Nie *et al.*, 2024d]	31.12	11.83	21.93	23.41
DDP (Ours)	**32.16**	**6.68**	**23.73**	**21.34**

Table 1: Quantitative results on the UCF Sports ($4 \rightarrow 6$ frames). ↑ / ↓ indicates the higher/lower values denote the better performance.

Unlike optical flow which only maps the current frame, this warping operation integrates all prior visual features $\mathbf{X}_t$ via the motion flow $\mathbf{F}_{t \rightarrow T}$, as shown in Fig. 6(b). This process takes place at each scale, and then the decoder produces future frames by converting the warped features from latent to pixel space.

4 Experiments

We demonstrate the effectiveness of the DDP model with multi-scenario evaluations. These scenarios are crucial for numerous applications requiring robust spatiotemporal predictive models. We demonstrate that the DDP model performs favorably against the state-of-the-art models on five challenging datasets corresponding to these scenarios.

Implementation Details. Our method uses PyTorch on an NVIDIA A100 GPU, training with 16-sequence mini-batches, the Adam optimizer, and the OneCycle scheduler. We apply a weight decay of $5e^{-2}$ and select learning rates from $\{1e^{-2}, 5e^{-3}, 1e^{-3}\}$ for stability. We use the MSE loss to supervise training and stochastic depth for regularization.

4.1 Human Motions: UCF Sports

Dataset and Setup. UCF Sports [Rodriguez *et al.*, 2008] comprises 150 videos from diverse sports scenes, depicting 10 distinct actions with complex human motion patterns. Following STRPM [Chang *et al.*, 2022], we scale resolution from 480×720 to 512×512, using 6,288 sequences for training and 752 for testing. The model observes 4 frames and predicts 1 frame ($4 \rightarrow 1$) during training and 6 frames ($4 \rightarrow 6$) during testing.

Main Results. Tab. 1 presents model performance, reporting PSNR and LPIPS metrics for $(T+1)_{th}$ and $(T+6)_{th}$ frames. DDP demonstrates significant performance gains over other methods. This dataset includes complex scenarios and motion patterns, like camera movement and motion blur. Our design effectively addresses these issues by separating visuals and motion, as shown in Fig. 2, demonstrating the potential for real-world applications and scalability to high-resolution spatiotemporal data.

4.2 Synthetic Motions: Moving MNIST

Dataset and Setup. The Moving MNIST [Srivastava *et al.*, 2015] dataset is constructed by randomly sampling two digits with 64×64 pixels from the MNIST dataset and making them float and bounce at boundaries with a constant direction and velocity. There are 10,000 sequences for training and 10,000 for testing. The model observes the first 10 frames and predicts the next 10 frames.

Main Results. Tab. 2 shows MSE and PSNR metrics of DDP network against the state-of-the-art predictive learning methods. Our model significantly outperforms these methods in both metrics. We also show a prediction example in Fig. 7(b). Notably, SimVP [Gao *et al.*, 2022] improves visual details through introducing IncepU, but missing part of the motion modeling leads to error accumulation over time as shown in the last row of Fig. 7(b). In contrast, DDP using a motion diffuser to model long-term motion patterns explicitly can mitigate this issue. This suggests that the DDP models synthetic motions better than other methods.

4.3 Driving Scenes: KITTI&Caltech

Dataset and Setup. The generalization ability is crucial for real-world driving scenes. The KITTI&Caltech [Geiger *et al.*, 2013; Dollár *et al.*, 2009] dataset evaluates generalization ability across different datasets. Following standard practice [Gao *et al.*, 2022], we train the model on the KITTI [Geiger *et al.*, 2013] dataset and evaluate it against the Caltech Pedestrian [Dollár *et al.*, 2009] dataset. We resize the resolution to 128×160, and models predict the next frame by previously observed 10 frames.

Method	MSE↓	PSNR↑
ConvLSTM [Shi *et al.*, 2015]	103.3	16.17
PredRNN [Wang *et al.*, 2017]	56.8	19.12
PredRNN++ [Wang *et al.*, 2018b]	46.5	20.11
Conv-TT-LSTM [Su *et al.*, 2020]	53.0	19.41
PredRNNv2 [Wang *et al.*, 2022]	48.4	20.12
SimVP [Gao *et al.*, 2022]	23.8	23.19
MMVP [Zhong *et al.*, 2023]	22.2	23.62
ModeRNN [Yao *et al.*, 2023]	42.1	20.45
WaST [Nie *et al.*, 2024d]	21.1	23.85
DDP (Ours)	**19.2**	**24.63**

Table 2: Quantitative results on the Moving MNIST ($10 \rightarrow 10$ frames) dataset.

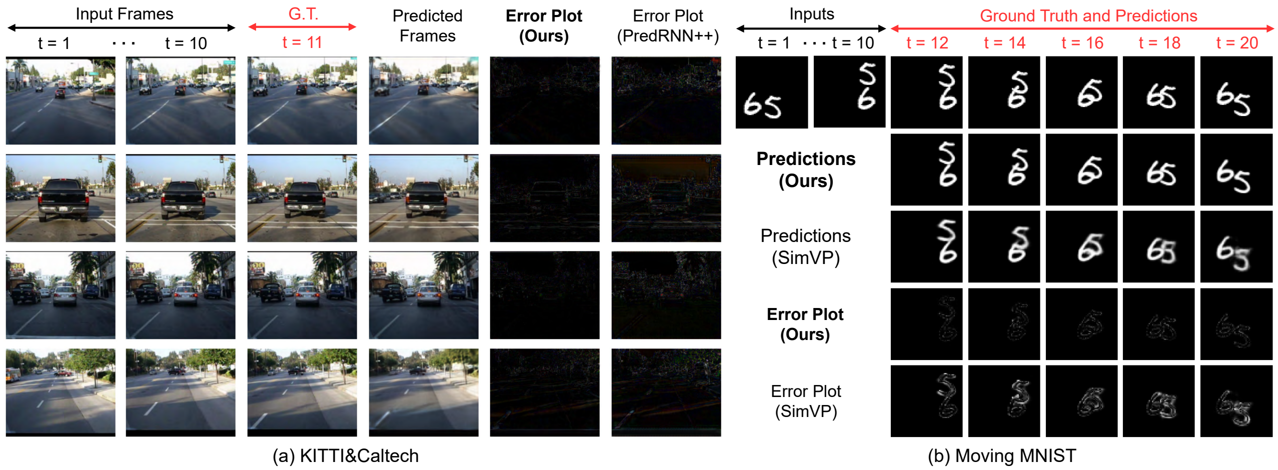

Figure 7: Qualitative results on the KITTI&Caltech ($10 \rightarrow 1$) and Moving MNIST ($10 \rightarrow 10$) datasets, where error plot = |ground truth − prediction| denotes the differences between the ground truth frames and their corresponding predicted frames.

Main Results. Tab. 3 shows the quantitative results of the proposed model and mainstream methods. The DDP model achieves strong performance under all metrics consistent with previous observations. These empirical results demonstrate the effectiveness of the DDP model for modeling spatiotemporal driving data. Qualitative visualizations Fig. 7(a) shows that our method can better predict lane lines and pinpoint distant, small entities than other methods, which indicates the potential generalization across scenes.

4.4 Traffic Flow: TaxiBJ

Dataset and Setup. TaxiBJ [Zhang and others, 2017] comprises taxi GPS trajectory data in Beijing, with 30-minute intervals and 32×32 spatial granularity. Models predict 4 future frames based on 4 observed frames. Complex road network dependencies and non-linear temporal dynamics have historically challenged traffic forecasting methods.

Main Results. Quantitative results are presented in Tab.3, with qualitative visualizations in Fig.8. TAU [Tan *et al.*, 2023], despite introducing a temporal attention unit and setting benchmarks in several datasets, fails to adequately capture road spatiotemporal dependencies, resulting in prediction inaccuracies (Fig. 8, last two rows). For unstructured data (*e.g.*, traffic, climate), intensity is similar to visual features in structured data. DDP still works well in both intensity and dynamics, consistently outperforming other approaches with minimal intensity differences across most regions. Notably, the optical flow-based DMVFN [Hu *et al.*, 2023] method underperforms in these scenes. In contrast, DDP, making no data-specific assumptions, demonstrates broader applicability across various modalities.

4.5 Global Climate: WeatherBench

Dataset and Setup. WeatherBench [Rasp *et al.*, 2020] contains climatic data from 1979 to 2018, re-gridded to $5.625°$ (32×64 grid points) and $1.40625°$ (128×256 grid points). We evaluate temperature prediction at $5.625°$ resolution, using 2010-2015 for training, 2016 for validation, and 2017-2018 for testing. The model forecasts 12-hour temperature based on 12-hour historical global temperature data.

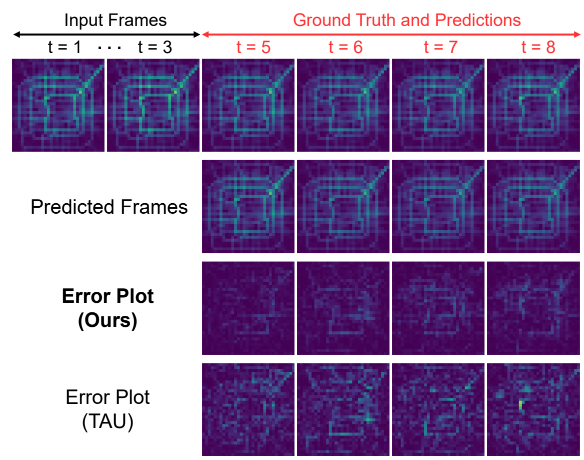

Figure 8: Qualitative results on the TaxiBJ ($4 \rightarrow 4$), where DDP captures road spatiotemporal dependencies

Main Results. For the WeatherBench dataset, Tab. 3 presents quantitative comparisons with state-of-the-art methods. DDP consistently outperforms all other state-of-the-art methods across all reported error metrics (MSE, MAE, and RMSE). It is particularly noteworthy that some methods, such as PhyDNet and DMVFN, exhibit significantly higher error values on this task, indicating substantial deviations in their predictions from the ground truth climate patterns. This comprehensive superiority suggests DDP's architecture is particularly well-suited for capturing the complex, long-range spatiotemporal dependencies inherent in climate data.

4.6 Ablation Studies

In this section, we further perform extensive ablation studies to study the components' effectiveness in our DDP.

Ablation of the Probabilistic Path. We implemented various spatiotemporal modules as motion diffusers (Tab. 4(a)),

Method	KITTI&Caltech			TaxiBJ			WeatherBench		
	MSE↓	MAE↓	SSIM↑	MSE↓	MAE↓	SSIM↑	MSE↓	MAE↓	RMSE↓
ConvLSTM [Shi *et al.*, 2015]	139.6	1583.3	0.9345	0.485	17.7	0.978	1.521	0.7949	1.233
PredRNN [Wang *et al.*, 2017]	130.4	1525.5	0.9374	0.464	17.1	0.971	1.331	0.7246	1.154
PredRNN++ [Wang *et al.*, 2018b]	129.6	1507.7	0.9453	0.448	16.9	0.977	1.634	0.7883	1.278
E3D-LSTM [Wang *et al.*, 2018c]	200.6	1946.2	0.9047	0.432	16.9	0.979	1.592	0.8059	1.262
PhyDNet [Guen and others, 2020]	312.2	2754.8	0.8615	0.419	16.2	0.982	285.9	8.7370	16.91
PredRNNv2 [Wang *et al.*, 2022]	147.8	1610.5	0.9330	0.383	15.6	0.983	1.545	0.7986	1.243
SimVP [Gao *et al.*, 2022]	160.2	1690.8	0.9338	0.414	16.2	0.982	1.238	0.7037	1.113
DMVFN [Hu *et al.*, 2023]	183.9	1531.1	0.9314	3.395	45.5	0.832	448.5	16.880	21.14
TAU [Tan *et al.*, 2023]	131.1	1507.8	0.9456	0.344	15.6	0.983	1.224	0.6810	1.106
SimVPv2 [Tan *et al.*, 2025]	129.7	1507.7	0.9454	0.324	15.0	0.984	1.105	0.6567	1.051
DDP (Ours)	**123.7**	**1418.9**	**0.9468**	**0.302**	**14.9**	**0.984**	**1.082**	**0.6332**	**1.040**

Table 3: Quantitative results of state-of-the-art methods on the KITTI&Caltech (10 → 1 frames), TaxiBJ (4 → 4 frames), and WeatherBench (12 → 12 frames) datasets. ↑ / ↓ indicates the higher/lower values denote the better performance.

utilizing convolution (Conv) and self-attention (SA). Replacing spatiotemporal SSMs with these variants, our experiments demonstrate that SSMs outperform other methods in terms of MSE metric, validating their efficacy in modeling spatiotemporal data. Moreover, We experimented with various backward scanning methods (Tab. 4(b)), where the spatiotemporal flip scanning yielded the lowest MSE.

(a) Motion modeling

Method	MSE↓
3D Conv	21.15
Spatiotemporal SA	20.96
Spatiotemporal SSMs	**19.23**

(b) Reverse scanning

Method	MSE↓
Spatial flip	23.39
Temporal flip	21.14
Spatiotemporal flip	**19.37**

Table 4: Ablation results of the probabilistic path on the Moving MNIST dataset.

Ablation of the Deterministic Path. We substituted the deterministic branch of the spectral-aware module (SAM) with various metaformer modules: Vision Transformer (ViT), Swin Transformer, and ConvNext, as shown in Fig. 9. Additionally, we compared different spectral architectures, including Fast Fourier Convolution (FFC), Fourier Neural Operator (FNO), and Wavelet Gating Network (WGN). Results demonstrate that our SAM, synthesizing the strengths of these frameworks, achieves optimal performance.

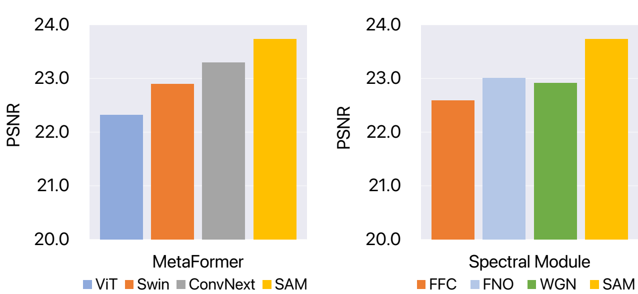

Figure 9: Ablation results of the deterministic path on the UCF Sports dataset.

Ablation of the Long-Term Prediction. To investigate the components affecting long-term predictions in DDP, we tested long-term input and output scenarios (Fig. 10). We replaced the Motion Diffuser (MDiff) with a deterministic model and restricted Motion-Visual Warping (MVW) to map only current frame latent features. Results indicate that the motion diffuser significantly impacts long-term dynamics, while MVW marginally enhances this capability.

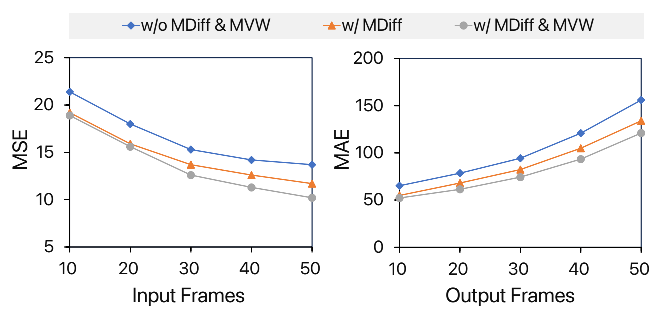

Figure 10: Ablation results of the long-term prediction on the Moving MNIST dataset.

5 Conclusion and Future Work

This paper presented Disentangling Deterministic and Probabilistic (DDP) modeling, a novel predictive learning framework employing two distinct latent pathways. The probabilistic pathway explicitly models complex, many-to-many motion patterns via a motion diffuser, while the deterministic pathway utilizes a spectral-aware enhancer to preserve and amplify visual details in the frequency domain. This dual-architecture design effectively balances robust visual consistency with the accurate capture of intricate, long-term motion dynamics. Comprehensive experimental validation demonstrates DDP's significant outperformance of existing state-of-the-art methods, yielding qualitatively superior visual predictions. Future research will focus on addressing the scalability to substantially longer sequences (*e.g.*, >100 frames) and higher resolutions (*e.g.*, 4K). These advancements present considerable challenges, requiring optimized resource management and highly efficient inference strategies.

Contribution Statement

Xuesong Nie and Haoyuan Jin made equal contributions.

References

[Bei *et al.*, 2021] Xinzhu Bei, Yanchao Yang, and Stefano Soatto. Learning semantic-aware dynamics for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 902–912, 2021.

[Chang *et al.*, 2022] Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. Strpm: A spatiotemporal residual predictive model for high-resolution video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13946–13955, 2022.

[Ding *et al.*, 2023] Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition. *arXiv preprint arXiv:2311.15599*, 2023.

[Dollár *et al.*, 2009] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE conference on computer vision and pattern recognition*, pages 304–311. IEEE, 2009.

[Gao *et al.*, 2022] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3170–3180, 2022.

[Geiger *et al.*, 2013] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[Gu and Dao, 2023] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[Guen and others, 2020] Vincent Le Guen et al. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11474–11484, 2020.

[Gupta *et al.*, 2022] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *arXiv preprint arXiv:2206.11894*, 2022.

[Hochreiter and others, 1997] Sepp Hochreiter et al. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Hu *et al.*, 2023] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2023.

[Islam and Bertasius, 2022] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022.

[Islam *et al.*, 2023] Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, and Gedas Bertasius. Efficient movie scene detection using state-space transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18749–18758, 2023.

[Jin *et al.*, 2024] Haoyuan Jin, Xuesong Nie, Yunfeng Yan, Xi Chen, Zhihang Zhu, and Donglian Qi. Object-level pseudo-3d lifting for distance-aware tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8015–8023, 2024.

[Karras *et al.*, 2022] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

[Lam *et al.*, 2022] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*, 2022.

[Liu *et al.*, 2024] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[Nguyen *et al.*, 2022] Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, Tri Dao, Stephen Baccus, and Christopher Ré. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Advances in neural information processing systems*, 35:2846–2861, 2022.

[Nie *et al.*, 2024a] Xuesong Nie, Xi Chen, Haoyuan Jin, Zhihang Zhu, Yunfeng Yan, and Donglian Qi. Triplet attention transformer for spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7036–7045, 2024.

[Nie *et al.*, 2024b] Xuesong Nie, Haoyuan Jin, Yunfeng Yan, Xi Chen, Zhihang Zhu, and Donglian Qi. Predtoken: Predicting unknown tokens and beyond with coarse-to-fine iterative decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18143–18152, 2024.

[Nie *et al.*, 2024c] Xuesong Nie, Haoyuan Jin, Yunfeng Yan, Xi Chen, Zhihang Zhu, and Donglian Qi. Scopevit: Scale-aware vision transformer. *Pattern Recognition*, 153:110470, 2024.

[Nie *et al.*, 2024d] 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*, volume 38, pages 4334–4342, 2024.

[Rasp *et al.*, 2020] 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*, 12(11):e2020MS002203, 2020.

[Rodriguez *et al.*, 2008] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[Shi *et al.*, 2015] 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*, 28, 2015.

[Srivastava *et al.*, 2015] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR, 2015.

[Su *et al.*, 2020] Jiahao Su, Wonmin Byeon, Jean Kossaifi, Furong Huang, Jan Kautz, and Anima Anandkumar. Convolutional tensor-train lstm for spatio-temporal learning. *Advances in Neural Information Processing Systems*, 33:13714–13726, 2020.

[Tan *et al.*, 2023] 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*, pages 18770–18782, 2023.

[Tan *et al.*, 2025] Cheng Tan, Zhangyang Gao, Siyuan Li, and Stan Z Li. Simvpv2: Towards simple yet powerful spatiotemporal predictive learning. *IEEE Transactions on Multimedia*, 2025.

[Villegas *et al.*, 2017] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017.

[Wang *et al.*, 2017] 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*, 30, 2017.

[Wang *et al.*, 2018a] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[Wang *et al.*, 2018b] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and S Yu Philip. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *International Conference on Machine Learning*, pages 5123–5132. PMLR, 2018.

[Wang *et al.*, 2018c] 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.

[Wang *et al.*, 2022] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, S Yu Philip, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2208–2225, 2022.

[Wang *et al.*, 2023] Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397, 2023.

[Wu *et al.*, 2020] Yue Wu, Rongrong Gao, Jaesik Park, and Qifeng Chen. Future video synthesis with object motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5539–5548, 2020.

[Wu *et al.*, 2021] Haixu Wu, Zhiyu Yao, Jianmin Wang, and Mingsheng Long. Motionrnn: A flexible model for video prediction with spacetime-varying motions, 2021.

[Yan *et al.*, 2023] Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. Diffusion models without attention. *arXiv preprint arXiv:2311.18257*, 2023.

[Yang *et al.*, 2021] Lingxiao Yang, Ru-Yuan Zhang, Lida Li, and Xiaohua Xie. Simam: A simple, parameter-free attention module for convolutional neural networks. In *International conference on machine learning*, pages 11863–11874. PMLR, 2021.

[Yao *et al.*, 2023] Zhiyu Yao, Yunbo Wang, Haixu Wu, Jianmin Wang, and Mingsheng Long. Modernn: Harnessing spatiotemporal mode collapse in unsupervised predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Yu *et al.*, 2022] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022.

[Zhang and others, 2017] Junbo Zhang et al. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[Zhong *et al.*, 2023] Yiqi Zhong, Luming Liang, Ilya Zharkov, and Ulrich Neumann. Mmvp: Motion-matrix-based video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4273–4283, 2023.

[Zhu *et al.*, 2024] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.