Interpretable and Generalizable Spatiotemporal Predictive Learning with Disentangled Consistency

Jingxuan Wei[1,2], Cheng Tan[3,4], Zhangyang Gao[3,4], Linzhuang Sun[1,2], Bihui Yu[1,2(✉)], Ruifeng Guo[1,2(✉)], and Stan Li[3,4(✉)]

[1] Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
weijingxuan20@mails.ucas.edu.cn, sunlinzhuang21@mails.ucas.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
{yubihui,grf}@sict.ac.cn
[3] Zhejiang University, Hangzhou, China
[4] AI Lab, Research Center for Industries of the Future, Westlake University, Hangzhou, China
{tancheng,gaozhangyang,stan.zq.li}@westlake.edu.cn

Abstract. In recent years, significant strides have been made in the field of spatiotemporal predictive learning, a discipline that focuses on accurately forecasting future sequences based on previously observed frames. Despite the impressive capabilities of current leading-edge models, which leverage specialized network architectures to optimize learning in both spatial and temporal domains, these models often fall short in their ability to accurately interpret underlying spatiotemporal dependencies and extend their learnings to unseen data. In this study, we attempt to address these shortcomings by disentangling the context and motion within sequential spatiotemporal data, and then systematically analyzing the relationship between the original and disentangled data. We introduce context-motion disentanglement modules that utilize temporal entropy to segregate the context and motion, and then apply regularization to the disentangled motion to ensure its consistency with the predicted frames produced by conventional spatiotemporal predictive learning. Our proposed methodology can be trained in an end-to-end fashion and serves to improve not just the predictive performance but also the interpretability and generalizability of the model. The efficacy of our proposed method is illustrated through comprehensive quantitative and qualitative assessments.

Keywords: Spatiotemporal predictive learning · self-supervised learning · convolutional neural networks · computer vision applications

1 Introduction

Deep learning has demonstrated considerable success in numerous domains [4,24–26,43,44,54]. A critical subfield of deep learning is spatiotemporal pre-

J. Wei and C. Tan—Equal Contribution.

A. Bifet et al. (Eds.): ECML PKDD 2024, LNAI 14943, pp. 3–20, 2024.
https://doi.org/10.1007/978-3-031-70352-2_1

dictive learning, a self-supervised learning discipline that focuses on forecasting future frames based on past observations. Previous studies have made commendable contributions by developing specialized modules to capture spatial correlations and temporal dependencies based on LSTM [16] and GRU [7]. Though these seminal works have achieved superior results, they face challenges in effectively interpreting the underlying spatiotemporal dependencies and generalizing the insights from disentangled information.

Past research [17,47] have strived to separate static contexts from dynamic motions, aiming to extract meaningful representations from sequential video data. The primary premise of these studies is that once the model successfully disentangles the context from the motion, it would have effectively learned the spatial correlations and temporal dependencies. Thus, they either build dual networks to separately capture motions and semantic contexts [11] or impose constraints in the latent spaces [17]. However, mediately predicting future frames by fusing the representations of contexts and motions usually performs worse than those directly optimizing for the future frames [12,52]. The reason to blame may be brute-force disentangling that destroys nonlinear spatiotemporal relations. Moreover, these methods employ disentangling in the latent space, which is difficult to present the actual disentangled contexts and motions explicitly. Their inherent complex architectures even hinder their interpretable ability.

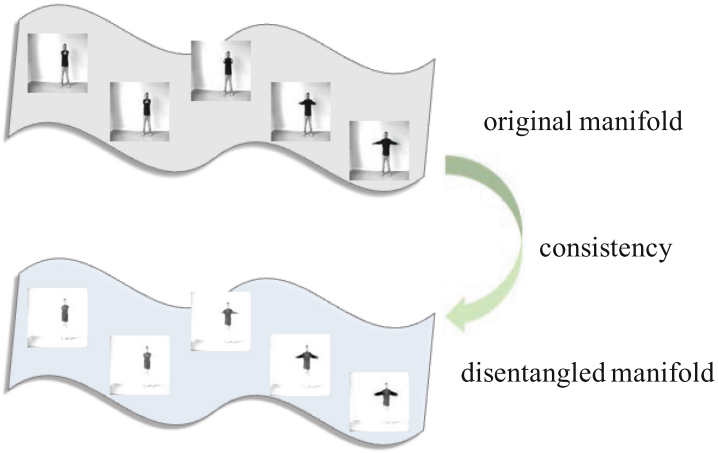


Fig. 1. The consistency between the manifolds of original sequential video data and disentangled representations.

Our study aspires to bridge this gap by fusing standard spatiotemporal learning with disentangled context-motion, creating a framework for interpretable and generalizable spatiotemporal learning. We introduce context-motion disentanglement modules leveraging temporal entropy to separate the context and motion. Based on the principles of manifold learning [27], we hypothesize that the original data and disentangled representations exist on different manifolds with analogous topological spaces. The assumption primarily comes from the basis of the static context and the dynamic motions. While the context is static, we can

regard the context as a constant that is added to the motion. We obtain the disentangled manifold from the original manifold minus a constant so that the manifolds are homeomorphic. As shown in Fig. 1, the disentangled representation containing varying motions should have similar spatiotemporal dependencies to the original data. By imposing a consistency constraint between manifolds, we exploit the disentangled representations in enhancing interpretable and generalizable spatiotemporal predictive learning.

2 Related Works

2.1 Spatiotemporal Predictive Learning

Recent advances in recurrent models [13,30] have provided valuable insights into spatiotemporal predictive learning [1,8,35,41,42,58]. Inspired by recurrent neural networks, VideoModeling [31] adopts language modeling and quantizes the image patches into an extensive dictionary for recurrent units. CompositeLSTM [39] further introduces the LSTM architecture and improves its performance. ConvLSTM [37] leverages convolutional neural networks to model the LSTM architecture. PredNet [29] continually predicts future video frames using deep recurrent convolutional neural networks with bottom-up and top-down connections. PredRNN [50] proposes a Spatiotemporal LSTM unit that simultaneously extracts and memorizes spatial and temporal representations. Its subsequential work PredRNN++ [52] further proposes a gradient highway unit and Casual LSTM adaptively capture temporal dependencies. E3D-LSTM [51] designs eidetic memory transition in recurrent convolutional units. Conv-TT-LSTM [40] employs a higher-order ConvLSTM to predict by combining convolutional features across time. MotionRNN [55] focuses on motion trends and transient variations. LMC-Memory [23] introduces a long-term motion context memory using memory alignment learning. PredCNN [57] and TrajectoryCNN [28] implement convolutional neural networks as the temporal module. SimVP [12] is a seminal work that applies Inception modules with a UNet architecture to learn the temporal evolution. TAU [45] proposes an attention-based temporal module that performs both intra-frame and inter-frame attention for spatiotemporal predictive learning.

2.2 Disentangled Representation

Decomposing the raw sequential video data into disentangled representations is an essential topic in the computer vision. DRNet [11] and MCnet [49] are early works on learning disentangled image representations from video. Their proposed methods aim to learn contexts and motions by two individual networks separately and then fuse the learned static and dynamic features in the latent space. MoCoGAN [47] shares a similar idea but generates video frames conditioned on random vectors. DDPAE [17] performs the video decomposition with multiple objects in addition to disentanglement and designs a specialized framework for

Moving MNIST. MGP-VAE [3] also models the latent space for disentangled representations in video sequences. While the previous studies focus on learning in the latent space, our method aims to explicitly present interpretable and generalizable spatiotemporal predictive learning by a disentangled consistency constraint.

3 Methods

3.1 Preliminaries

We formally define the spatiotemporal predictive learning problem as follows. Given a video sequence $\boldsymbol{X}^{t,T} = \{\boldsymbol{x}^i\}_{t-T+1}^{t}$ at time t with the past T frames, we aim to predict the subsequent T' frames $\boldsymbol{Y}^{t+1,T'} = \{\boldsymbol{x}^i\}_{t+1}^{t+T'}$ from time $t+1$, where $\boldsymbol{x}^i \in \mathbb{R}^{C\times H\times W}$ is usually an image with channels C, height H, and width W. In practice, we represent the video sequences as tensors, i.e., $\boldsymbol{X}^{t,T} \in \mathbb{R}^{T\times C\times H\times W}$ and $\boldsymbol{Y}^{t+1,T'} \in \mathbb{R}^{T'\times C\times H\times W}$.

The model with learnable parameters Θ learns a mapping $\mathcal{F}_\Theta : \boldsymbol{X}^{t,T} \mapsto \boldsymbol{Y}^{t+1,T'}$ by exploring both spatial and temporal dependencies. In our case, the mapping $\mathcal{F}_\Theta$ is a neural network model trained to minimize the difference between the predicted future frames and the ground-truth future frames. The optimal parameters Θ^* are:

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

where $\mathcal{L}$ is a loss function that evaluates such differences. By optimizing such a loss function, the model is able to learn the inherent spatiotemporal dependencies and thus accurately predicts future frames.

We recognize context and motion as semantically static and dynamic objects, respectively. The data $\boldsymbol{X}$ are assumed to consist of the context $\boldsymbol{c} \in \mathbb{R}^{C\times H\times W}$ and the motion $\boldsymbol{O} = \{\boldsymbol{o}_i | \boldsymbol{o}_i \in \mathbb{R}^{C\times H\times W}\}$. The context and the motion are controlled by the state of movement $\boldsymbol{S} = \{\boldsymbol{s}_i | \boldsymbol{s}_i \in \mathbb{R}^{1\times H\times W}\}$. For each frame $\boldsymbol{x}^i$ in $\mathcal{X}$, the formal representation is:

$$\boldsymbol{x}^i = \boldsymbol{o}^i \odot \boldsymbol{s} + \boldsymbol{c} \odot (1 - \boldsymbol{s}), \forall \boldsymbol{x}_i \in \boldsymbol{X}, \boldsymbol{o}_i \in \boldsymbol{O}, \boldsymbol{s}_i \in \boldsymbol{S}, \tag{2}$$

where $\odot$ is the Hadamard product.

In this study, we decouple the context and motion of each frame through explicit context-motion disentanglement mechanism and implicit disentangled consistency for presenting an interpretable and generalizable spatiotemporal predictive learning.

3.2 Context-Motion Disentanglement

We first decompose the desired mapping $\mathcal{F}$ into two submappings:

$$\mathcal{F} \triangleq \mathcal{H} \circ \mathcal{G}, \tag{3}$$

where $\mathcal{H} : \boldsymbol{X}^{t,T} \mapsto \boldsymbol{H}^t$, $\mathcal{G} : \boldsymbol{H}^t \mapsto \boldsymbol{Y}^{t+1,T'}$, and $\boldsymbol{H}^t \in \mathbb{R}^{T' \times C \times H \times H \times W}$ is the latent representation at time t that contains information from previous T and following T'. $\mathcal{H}$ can be an arbitrary mapping that aims to explore the underlying spatiotemporal dependencies of the input frames $\mathcal{X}^{t,T}$ and project it into an informative latent space. In contrast to the mapping $\mathcal{H}$, the latter mapping $\mathcal{G}$ reconstructs the visual imaging and predicts the future frames $\boldsymbol{Y}^{t+1,T'}$ based on the representation $\boldsymbol{H}^t$ in the latent space.

For standard spatiotemporal predictive learning methods, $\mathcal{G}$ can be an arbitrary mapping, as well as $\mathcal{H}$. In this study, we explicitly define the mapping $\mathcal{G}$ for specific context-motion disentanglement:

$$\mathcal{G} \triangleq \boldsymbol{O} \odot \boldsymbol{S} + \boldsymbol{c} \odot (1 - \boldsymbol{S}), \tag{4}$$

where we practically represent the sets as tensors, i.e., $\boldsymbol{O} \in \mathbb{R}^{T' \times C \times H \times W}$ and $\boldsymbol{S} \in \mathbb{R}^{T' \times 1 \times H \times W}$. The context tensor $\boldsymbol{c} \in \mathbb{R}^{1 \times C \times H \times W}$ is a tensor variation compared to the definition in Sect. 3.1. The motion tensor $\boldsymbol{O}$, context tensor $\boldsymbol{c}$, and state tensor $\boldsymbol{S}$ are obtained by mappings $\mathcal{O} : \boldsymbol{H} \mapsto \boldsymbol{O}$, $\mathcal{C} : \boldsymbol{H} \mapsto \boldsymbol{c}$, and $\mathcal{S} : \boldsymbol{H} \mapsto \boldsymbol{S}$, respectively.

Though the $\mathcal{G}$ is specified to decouple the context and motion, directly optimizing the mean square error (MSE) loss alone, as standard spatiotemporal predictive learning does, is unreliable. The MSE loss cannot guide the neural network automatically separate the context and motion. We argue that the key to context-motion disentanglement is to determine the context accurately. Thus, we impose the inductive bias that the pixels in context are likely to be static across the varying time.

To evaluate the inherent uncertainty of video frames, we intuitively borrow the concept of entropy from information theory. Here we refer to $\Delta \boldsymbol{x}^i$ as a pixel in a specific position of frame $\boldsymbol{x}^i$ and $\Delta \boldsymbol{X}$ as the pixel in the same position of all frames in $\boldsymbol{X}$. We define the probability of whether this pixel is changing Δw^i as:

$$\Delta w^i = \frac{\Delta \boldsymbol{x}^i - \Delta \boldsymbol{x}^0}{\max \Delta \boldsymbol{x} - \min \Delta \boldsymbol{x}}, \tag{5}$$

which is normalized in $[0, 1]$ according to the changing scope compared to the initial frame. The uncertainty of whether the pixel belongs to the context is evaluated by its average entropy of w:

$$E(\Delta w) = -\frac{1}{T} \sum_{i=t-T+1}^{t} p(\Delta w_i) \log p(\Delta w_i), \tag{6}$$

then we obtain a mask $\boldsymbol{M} \in \{0, 1\}^{1 \times C \times H \times W}$ that should be able to filter *reliable context* by a threshold $\bar{w}$. For each pixel, if the corresponding E is lower than $\bar{w}$, we recognize it as the static context, i.e., $\boldsymbol{M}$ has a value of 1 for this pixel and vice versa.

With the inductive bias of reliable context given by $\boldsymbol{M}$, we design the disentanglement loss as:

$$\mathcal{L}_d(\boldsymbol{X}) = \frac{1}{T'} \sum_{t+1}^{t+T'} \|(\boldsymbol{c} - \boldsymbol{x}^i) \odot \boldsymbol{M}\|. \tag{7}$$

This loss guarantees that at least the reliable static context is learned. Taking advantage of the flatness of convolutional networks, the model can disentangle actual context based on the above reliable context.

3.3 Disentangled Consistency

Despite the disentanglement loss $\mathcal{L}_d$ enforcing explicit model discrimination between context and motion, it remains reliant on the inductive bias $\boldsymbol{M}$. We contend that the context is intrinsically static in its semantics and that the disentangled frames should exhibit consistency with the actual frames. Consider a manifold $\mathcal{M}_x$ representative of the original data, with the correlated disentangled representations inhabiting another manifold, denoted as $\mathcal{M}_o$. s **Definition.** We define two topological spaces, denoted as $\mathcal{M}_x$ and $\mathcal{M}_o$, to be homeomorphic if and only if there exists a bijective mapping function $f : \mathcal{M}_x \mapsto \mathcal{M}_o$ with the following properties: (i) The function f is continuous. (ii) The inverse of f, denoted as f^{-1}, exists and is also continuous.

This definition [10,15,32,38] reveals the relationship between the manifold $\mathcal{M}_x$ and $\mathcal{M}_o$. According to Eq. 4, we can observe that once the mapping is bijective the disentangled manifold is homeomorphic to the original manifold. In other words, the original manifold $\mathcal{M}_x$ and the disentangled manifold $\mathcal{M}_o$ are topological equivalences.

Theorem. Given a homeomorphism $f(\boldsymbol{X})$, a mapping that is both smooth and possesses a unique inverse, the mutual information is invariant under such transformation, such that $I(\boldsymbol{X}, \boldsymbol{O}) = I(f(\boldsymbol{X}), \boldsymbol{O})$.

Proof. First, remember that the entropy of a discrete random variable $\boldsymbol{X}$ is defined as $H(\boldsymbol{X}) = -\sum_{x \in \boldsymbol{X}} p(x) \log p(x)$, where $p(x)$ is the probability mass function of X. For continuous random variables, the entropy is similarly defined but with an integral instead of a sum, and the probability density function instead of the probability mass function.

Now consider a homeomorphism f, and suppose $p_{\boldsymbol{X}}(x)$ is the probability density function of $\boldsymbol{X}$ and $p_{\boldsymbol{O}}(o)$ is the probability density function of $\boldsymbol{O}$, which equals to $p_{\boldsymbol{X}}(f^{-1}(o))$ due to the invariance of probability under the transformation.

The differential entropy $H(\boldsymbol{O})$ of $\boldsymbol{O}$ is then:

$$\begin{aligned} H(\boldsymbol{O}) &= -\int p_{\boldsymbol{O}}(o) \log p_{\boldsymbol{O}}(o) do \\ &= -\int p_{\boldsymbol{X}}(f^{-1}(o)) \log p_{\boldsymbol{X}}(f^{-1}(o)) do, \end{aligned} \tag{8}$$

By changing the variable from o to $x = f^{-1}(o)$, and remembering that homeomorphisms preserve the measure, the differential entropy $H(\mathbf{O})$ of $\mathbf{O}$ transforms to:

$$H(\boldsymbol{O}) = -\int p_{\boldsymbol{X}}(x) \log p_{\boldsymbol{X}}(x) dx = H(\boldsymbol{X}). \tag{9}$$

So, the entropy of $\mathbf{X}$ and $\mathbf{O}$ are equal. Since the entropy is invariant under homeomorphisms, the conditional entropy is also invariant. Therefore, mutual information, which is a combination of entropy and conditional entropy, is also invariant under homeomorphisms.

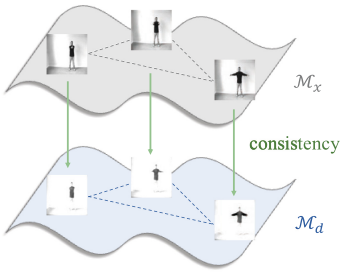


Fig. 2. Characterize the relationship between $\mathcal{M}_x$ and $\mathcal{M}_d$ from the geometric viewpoint and regularize the geometric property to be consistent.

This theorem [9,21,46] reveals the connections between $\mathcal{M}_x$ and $\mathcal{M}_o$. If the mapping f is bijective, their mutual information is:

$$I(\boldsymbol{X}, \boldsymbol{O}) = H(\boldsymbol{X}) + H(\boldsymbol{O}) - H(\boldsymbol{X}, \boldsymbol{O}) \tag{10}$$

is maximized. Based on the above observation, we characterize the relationship between the manifolds $\mathcal{M}_x$ and $\mathcal{M}_o$ from the geometric viewpoint. As shown in Fig. 2, we consider the pairwise distance as the key geometric property and regularize the manifold $\mathcal{M}_o$ to have a similar geometric structure as $\mathcal{M}_x$. For those limited data points, the mapping f is approaching bijective through preserving this geometric property.

Then, we define the pairwise distances in the two manifolds as follows:

$$d_x = \frac{\|\boldsymbol{x}^i - \boldsymbol{x}^j\|}{\sqrt{D}}, \; d_o = \frac{\|\boldsymbol{o}^i - \boldsymbol{o}^j\|}{\sqrt{D}}, \tag{11}$$

where $\|\cdot\|$ is Euclidean distance, $D = C \times H \times W$ is a scale factor for avoiding large magnitude [48], $i, j \in \{t+1, ..., t+T'\}$, and $i \neq j$. To model the distance in a nonlinear manner and obtain expressive metrics, we project the distance into normal distributions:

$$\begin{aligned} p(d_x) &= \frac{C_x}{\sigma_x \sqrt{2\pi}} \exp\big(-\frac{d_x^2}{2\sigma_x^2}\big), \\ p(d_o) &= \frac{C_o}{\sigma_o \sqrt{2\pi}} \exp\big(-\frac{d_o^2}{2\sigma_o^2}\big), \end{aligned} \tag{12}$$

where C_x, C_o are constants that forces the $p(\cdot) \in [0, 1]$, and σ_x, σ_o are controllable hyperparameters. For the convenience of optimization, we empirically assumes $p(d_g), p(d_o) \sim N(0, \frac{1}{2})$ in the experiments.

The disentangled consistency is formulated as:

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{X}, \boldsymbol{O}) = &- p(d_x) \log(p(d_o)) \\ &- (1 - p(d_x)) \log(1 - p(d_o)), \end{aligned} \tag{13}$$

in which the binary cross entropy between $p(d_x)$ and $p(d_o)$ is expected to be minimized.

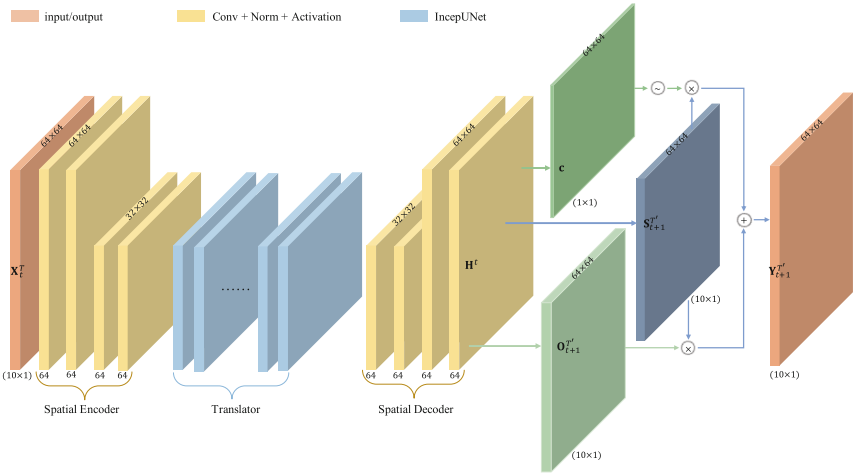


Fig. 3. The model architecture of our proposed method with the input from Moving MNIST. We employ a simple encoder-decoder model as the base architecture. The decoded representation $\boldsymbol{H}^t$ is used to obtain the context $\boldsymbol{c}$, motion $\boldsymbol{O}_{t+1}^{T'}$ and state $\boldsymbol{S}_{t+1}^{T'}$.

3.4 Practical Implementation

We implement our proposed method by modifying the network of the current state-of-the-art method SimVP [12]. SimVP is a solid baseline in spatiotemporal predictive learning. As shown in Fig. 3, a spatial encoder and a spatial decoder are simple convolutional networks with downsampling and upsampling operations, while a translator network is in the middle for learning the spatiotemporal correlations. In SimVP, the translator network consists of blocks of Inception-UNet (IncepUNet). We remove the last layer of SimVP and employ the output of the penultimate layer as $\boldsymbol{H}^t$. The mappings $\mathcal{O}, \mathcal{C}, \mathcal{S}$ are implemented by one-layer convolutional networks that project $\mathcal{H}$ to $\boldsymbol{O}_{t+1}^{T'}, \boldsymbol{S}_{t+1}^{T'}$, and $\boldsymbol{c}$, respectively.

The overall loss function is a linear combination of MSE loss, disentanglement loss $\mathcal{L}_d$, and disentangled consistency loss $\mathcal{L}_c$:

$$\begin{aligned}\mathcal{L} &= \text{MSE}(\mathcal{F}_\Theta(\boldsymbol{X}^{t,T}), \boldsymbol{Y}^{t+1,T'}) + \alpha\mathcal{L}_d + \beta\mathcal{L}_c, \\ &= \|\mathcal{F}_\Theta(\boldsymbol{X}^{t,T}) - \boldsymbol{Y}^{t+1,T'}\|^2 + \alpha\mathcal{L}_d + \beta\mathcal{L}_c,\end{aligned} \tag{14}$$

where α, β are weights of loss $\mathcal{L}_d$ and $\mathcal{L}_c$. We empirically set the values as $\alpha = 1.0, \beta = 0.1$ in default.

It is worth noting that though our proposed method is implemented based on the baseline SimVP, it is also suitable for other spatiotemporal predictive learning baselines.

4 Experiments

We evaluate our method by both quantitative and qualitative validation. We present the interpretability across different experimental settings as follows: (1) standard spatiotemporal predictive learning, (2) generalizing to unknown scenes.

4.1 Standard Spatiotemporal Predictive Learning

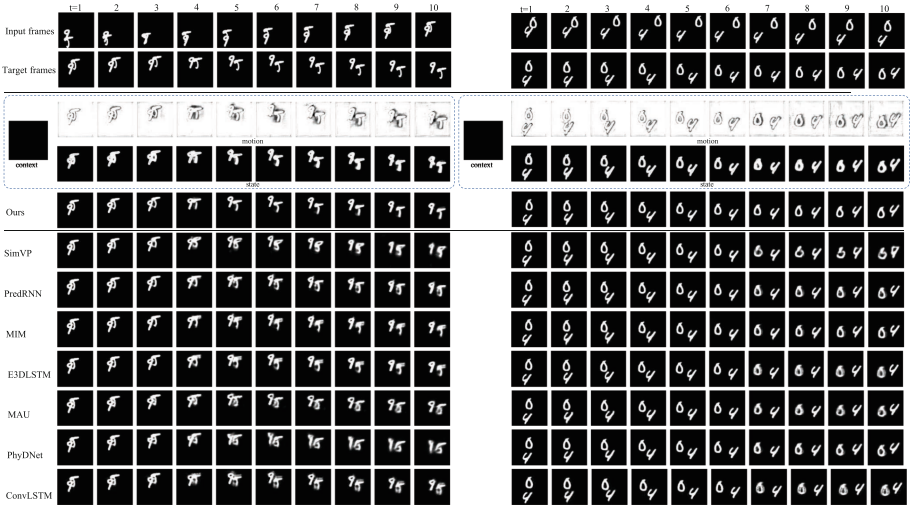


Fig. 4. Qualitative results on the Moving MNIST dataset. We show the disentangled context, motion, and state in the dotted boxes.

Moving MNIST. The Moving MNIST dataset [39], a widely recognized benchmark in standard spatiotemporal predictive learning, is a synthetic compilation. It comprises two individual digits meandering within a 64×64 grid, reacting to boundaries with a bounce-back motion. The task involves predicting the

subsequent 10 frames, given a historical sequence of 10 frames. Our proposed methodology addresses this by explicitly disentangling the complex spatiotemporal dependencies and capitalizing on the ensuing disentangled consistency for improved performance. It is anticipated that our model will demonstrate a high level of proficiency in predicting future frames with precision.

Our experimental setup parallels the one detailed in SimVP [12]. We measure our approach's performance against formidable benchmarks, including ConvLSTM [37], PredRNN [50], E3D-LSTM [51], MotionGRU [55], CrevNet [59], PhyDNet [14], SimVP [12], and others. We also compare our results with advanced techniques such as PhyDNet [14] and DDPAE [17], which engage in latent space disentanglement. The efficacy of our method is evidenced through quantitative metrics-frame-wise Mean Squared Error (MSE), Mean Absolute Error (MAE), and Structural Similarity Index Measure (SSIM)-and showcased in Table 1. To supplement our quantitative results, we offer a visual representation of our qualitative findings in Fig. 4. It becomes clear that our approach surpasses other state-of-the-art methods in performance, attributing its success to the robust modeling of context and motion. This capability confers our model with a competitive advantage, enabling it to outperform its counterparts.

Table 1. Quantitative results of different methods on the Moving MNIST dataset (10 → 10 frames).

Method	Moving MNIST (2 digits)		
	MSE↓	MAE↓	SSIM↑
ConvLSTM [37]	103.3	182.9	0.707
PredRNN [50]	56.8	126.1	0.867
PredRNN++ [52]	46.5	106.8	0.898
MIM [53]	44.2	101.1	0.910
LMC [23]	41.5	–	0.924
E3D-LSTM [51]	41.3	87.2	0.910
Conv-TT-LSTM [40]	53.0	–	0.915
DDPAE [17]	38.9	90.7	0.922
CrevNet [59]	38.5	–	0.928
MotionGRU [55]	34.3	–	0.928
CMS-LSTM [5]	33.6	73.1	0.931
MAU [6]	27.6	–	0.937
PhyDNet [14]	24.4	70.3	0.947
SimVP [12]	23.8	68.9	0.948
Ours	**22.9**	**68.6**	**0.949**

KTH. The KTH dataset [36], a compendium of human poses, encapsulates 25 individuals performing six distinct actions: walking, jogging, running, boxing, hand waving, and hand clapping. The intricacies of human motion stem from

the stochastic nature of various individuals performing different actions. The KTH dataset, however, is noted for its relatively consistent motion patterns. By studying historical frames, our model-built on the principles of interpretable and generalizable spatiotemporal predictive learning-is engineered to comprehend the dynamics of human motion and, subsequently, to anticipate long-term changes in future poses. Additionally, the prediction of extended sequences accurately is a nuanced problem in conventional spatiotemporal predictive learning. Our method strives to cultivate an interpretable and robust model, harnessing the learned spatiotemporal dependencies to predict long sequences with precision.

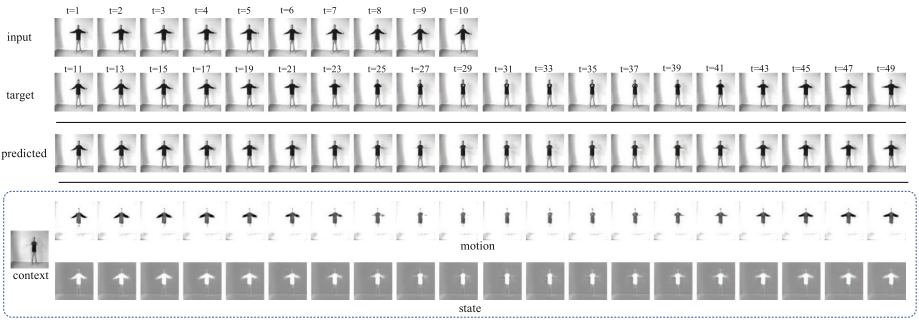


Fig. 5. Qualitative results on the KTH dataset. The example is predicting the next 40 frames based on the given historical 10 frames. The context $\boldsymbol{c}$, motion $\boldsymbol{O}$, and state $\boldsymbol{S}$ are shown in the dotted box.

Our experimental framework mirrors the one employed in SimVP [12], with the model being trained for 100 epochs. The evaluation of its performance is carried out using the SSIM and PSNR metrics [34,51]. Empirically, SSIM tends to focus more on disparities in visual sharpness, while PSNR leans towards pixel-level accuracy. By taking both these metrics into account, we ensure a comprehensive evaluation of the models. We compare the performance under two distinct settings: predicting the next 20 or 40 frames based on the historical ten frames. As depicted in Table 2, our method outperforms state-of-the-art methods on the KTH dataset in both the $10 \rightarrow 20$ and $10 \rightarrow 40$ scenarios. Despite the notable accomplishments of previous baselines, our method still demonstrates superior performance, thereby underscoring the efficacy of delving into context-motion disentanglement and implementing a disentangled consistency strategy.

We visualize an example of the predicted and disentangled results in Fig. 5. It can be seen that the model captures the static part that consists of a scene and a person with blurry arms as the context. The motion captures the details of the arms when it is swung. The result is controlled by the state that determines the proportion of dynamic and static parts. The motion ignores details of the scene and the static legs of the person but clearly delineates the swinging arms.

Table 2. Quantitative results of different methods on the KTH dataset (10 → 20 frames and 10 → 40 frames).

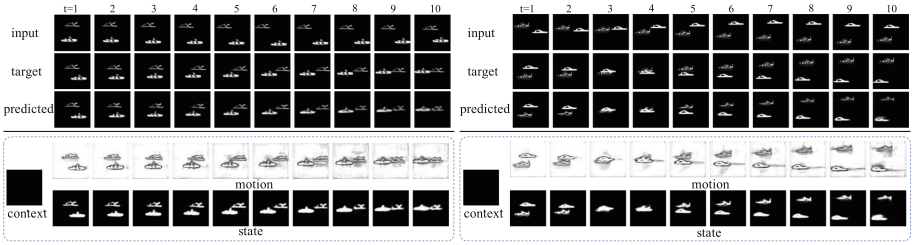
Method	KTH (10 → 20)		KTH (10 → 40)	
	SSIM↑	PSNR↑	SSIM↑	PSNR↑
MCnet [49]	0.804	25.95	0.73	23.89
ConvLSTM [37]	0.712	23.58	0.639	22.85
DFN [18]	0.794	27.26	0.652	23.01
fRNN [33]	0.771	26.12	0.678	23.77
SV2Pv [2]	0.838	27.79	0.789	26.12
PredRNN [50]	0.839	27.55	0.703	24.16
VarNet [19]	0.843	28.48	0.739	25.37
SAVP-VAE [22]	0.852	27.77	0.811	26.18
PredRNN++ [52]	0.865	28.47	0.741	25.21
E3d-LSTM [51]	0.879	29.31	0.810	27.24
STMFANet [20]	0.893	29.85	0.851	27.56
SimVP [12]	0.905	33.72	0.886	32.93
Ours	**0.909**	**33.97**	**0.890**	**33.20**

4.2 Generalizing to Unknown Scenes

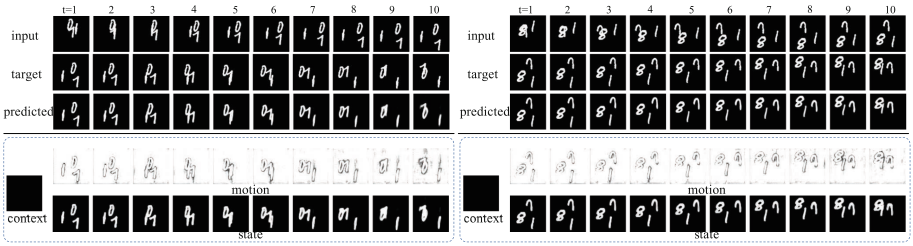
Unknown Object. Our method benefits from the robust modeling of spatiotemporal dependencies that exploits the relationship between context and motion in both explicit and implicit ways. To verify the robustness and generalization ability, we make the Moving Fashion MNIST dataset by replacing digits with objects of Fashion MNIST [56]. We use the models pre-trained on Moving MNIST to evaluate the performance on Moving Fashion MNIST.

Table 3. Quantitative results on the Moving Fashion MNIST dataset (10 → 10 frames).

Method	Moving Fashion MNIST (2 objects)		
	MSE↓	MAE↓	SSIM↑
ConvLSTM [37]	96.2	268.1	0.628
PredRNN [50]	90.6	225.6	0.749
PredRNN++ [52]	82.2	204.5	0.782
MIM [53]	80.6	204.7	0.778
E3D-LSTM [51]	78.3	196.9	0.791
PhyDNet [14]	86.7	207.7	0.774
MAU [6]	82.8	201.0	0.781
SimVP [12]	79.1	196.9	0.789
Ours	**72.3**	**178.5**	**0.810**



(a) Moving Fashion MNIST with two objects



(b) Moving MNIST with three objects

Fig. 6. Qualitative results on unknown scenes. With the model pretrained on the vanilla Moving MNIST dataset, we show the visualization of evaluating Moving Fashion MNIST and three-object Moving MNIST. The disentangled context, motion and state are shown in dotted boxes below the predicted frames.

Table 4. Quantitative results on the Moving MNIST dataset with three digits (10 → 10 frames).

Method	Moving MNIST (3 digits)		
	MSE↓	MAE↓	SSIM↑
ConvLSTM [37]	115.4	262.0	0.730
PredRNN [50]	101.8	235.7	0.766
PredRNN++ [52]	84.4	203.3	0.808
MIM [53]	83.1	198.0	0.816
E3D-LSTM [51]	75.6	186.0	0.829
PhyDNet [14]	85.6	179.2	0.833
MAU [6]	71.2	159.0	0.859
SimVP [12]	75.2	165.7	0.849
Ours	**68.9**	**144.1**	**0.870**

We show the quantitative results in Table 3. It can be seen that our model achieves significantly better performance than baseline models. Specifically, our model improves the state-of-the-art SimVP model by about 8.60% in the MSE metric and about 9.34% in the MAE metric, indicating the strong generalization ability of our model. The qualitative results in Fig. 6(a) show that our model captures context and motion well despite the objects have changed.

Unknown Setting. We extend our exploration to test the model's generalizability in a more complex setting that incorporates three moving digits instead of the customary two. This approach aligns with the strategy delineated in Sect. 4.2, wherein the model, initially trained on the Moving MNIST dataset with two digits, is employed to gauge its performance on a Moving MNIST variant with three digits. To put it differently, we train the model using data containing two digits and evaluate its performance with data featuring three digits. The model is expected to identify the dynamic elements, as opposed to merely recalling the previously observed scenarios.

As represented in Table 4, our model consistently surpasses baseline models by a considerable margin across all metrics. Specifically, our model enhances the state-of-the-art SimVP model by approximately 9.14% in the MSE metric and around 13.03% in the MAE metric. We illustrate a predicted example in Fig. 6(b). Although the context-motion disentanglement mechanism distinctly recognizes the dynamic and static elements, the predicted frames closely resemble the ground-truth frames. These experimental results affirm that our model, by learning the context-motion, exhibits a formidable generalization capacity.

4.3 Ablation Study

A series of ablation studies have been conducted on both Moving MNIST and Moving Fashion MNIST datasets, with the MSE metrics reported in Table 5. Initially, we eliminate the disentangled consistency, a mechanism that implicitly disentangles the context and motion. As a result, we observe a significant deterioration in performance on the Moving Fashion MNIST dataset, underlining the pivotal role disentangled consistency plays in enhancing the model's generalization capabilities. Subsequently, when we remove the context-motion disentanglement modules, the performance suffers an even more profound degradation.

Table 5. Ablation study of our proposed method. (MSE ↓)

Method	M-MNIST	M-FMNIST
Ours	22.9	72.3
w/o disentangled consistency	23.2	75.4
w/o disentanglement modules	23.8	79.1

5 Limitations

5.1 Reverse Problem

Our model is designed to forecast subsequent sequences given an input sequence $\{\boldsymbol{x}_i, \boldsymbol{x}_{i+1}, ..., \boldsymbol{x}_{i+T}\}$, yielding a prediction of the form $\{\boldsymbol{x}_{i+T}, \boldsymbol{x}_{i+T+1}, ..., \boldsymbol{x}_{i+T'}\}$.

An intriguing question that arises is whether the model could perform a "reverse prediction" if the input is rearranged as $\{\boldsymbol{x}_{i+T}, \boldsymbol{x}_{i+T-1}, ..., \boldsymbol{x}_i\}$, essentially predicting a sequence of the form $\{\boldsymbol{x}_{i-1}, \boldsymbol{x}_{i-2}, ..., \boldsymbol{x}_{i-T'}\}$. We refer to this scenario as the "reverse prediction problem". The potential of our model, which excels in interpretable and generalizable spatiotemporal predictive learning, to address this challenge presents a fascinating direction for future exploration. This innovative application could provide valuable insights into how prediction models can be utilized in more flexible and versatile ways.

5.2 Handling of Irregularly Sampled Data

The datasets utilized in this research were sampled at consistent time intervals. This approach may not be well-suited for handling irregularly sampled data, such as those with missing values. Our method may struggle to learn from such data and to make predictions for arbitrary timestamps in the future. A plausible solution to this challenge might involve appending timestamp information to the input or hidden feature vectors during the generation phase. Alternatively, neural ordinary differential equations could be employed to model time-continuous data.

5.3 Adaptability to Dynamic Views

The proposed context-motion disentanglement module is likely more compatible with static views, as it operates under the assumption that the background of the same video remains largely unchanged. The extension of our disentanglement strategy to more dynamic views represents an interesting area for future research.

6 Conclusion

In this work, we present an interpretable and generalizable spatiotemporal predictive learning method, which seeks to disentangle the context and the motion from sequential spatiotemporal data. Specifically, we design a context-motion disentanglement mechanism and a disentangled consistency strategy to perform both explicit and implicit context-motion disentanglement. Our experimental results demonstrate that our proposed model adeptly decouples static context from dynamic motion, and further learns the nuanced spatiotemporal dependencies, outshining models that merely rely on rote memorization. Crucially, our model demonstrates robust generalization to previously unseen data. We anticipate that the methodology we have put forth may provide fresh insights and potentially stimulate advancements in the sphere of artificial general intelligence.

Acknowledgements. This work is supported by the Shenyang Science and Technology Plan, grant number 23-407-3-29.

Ethical Statement
Our submission does not involve any ethical issues, including but not limited to privacy, security, etc.

References

1. Acharya, D., Huang, Z., Paudel, D.P., Van Gool, L.: Towards high resolution video generation with progressive growing of sliced wasserstein gans. arXiv preprint arXiv:1810.02419 (2018)
2. Babaeizadeh, M., et al.: Stochastic variational video prediction. In: ICLR (2018)
3. Bhagat, S., Uppal, S., Yin, Z., Lim, N.: Disentangling multiple features in video sequences using Gaussian processes in variational autoencoders. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12368, pp. 102–117. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58592-1_7
4. Cao, H., et al.: A survey on generative diffusion models. IEEE Trans. Knowl. Data Eng. (2024)
5. Chai, Z., et al.: CMS-LSTM: context embedding and multi-scale spatiotemporal expression LSTM for predictive learning. In: ICME, pp. 01–06 (2022)
6. Chang, Z., et al.: Mau: a motion-aware unit for video prediction and beyond. Adv. NIPS **34** (2021)
7. Cho, K., et al.: On the properties of neural machine translation: encoder–decoder approaches. In: Proceedings of SSST-8, pp. 103–111 (2014)
8. Clark, A., Donahue, J., Simonyan, K.: Adversarial video generation on complex datasets. arXiv preprint arXiv:1907.06571 (2019)
9. Cover, T.M., Thomas, J.A.: Elements of information theory second edition solutions to problems. Internet Access, 19–20 (2006)
10. Crossley, M.D.: Essential Topology. Springer, Heidelberg (2006). https://doi.org/10.1007/1-84628-194-6
11. Denton, E.L., et al.: Unsupervised learning of disentangled representations from video. Adv. NIPS **30** (2017)
12. Gao, Z., Tan, C., Li, S.Z.: Simvp: simpler yet better video prediction. In: Proceedings of CVPR, pp. 3170–3180 (2022)
13. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: International Conference on Machine Learning, pp. 1243–1252. PMLR (2017)
14. Guen, V.L., Thome, N.: Disentangling physical dynamics from unknown factors for unsupervised video prediction. In: Proceedings of CVPR, pp. 11474–11484 (2020)
15. Hawking, S.W., Ellis, G.F.R.: The Large Scale Structure of Space-Time, vol. 1. Cambridge University Press, Cambridge (1973)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
17. Hsieh, J.T., et al.: Learning to decompose and disentangle representations for video prediction. Adv. NIPS **31** (2018)
18. . Jia, X., et al.: Dynamic filter networks. Adv. NIPS **29** (2016)
19. Jin, B., et al.: Varnet: exploring variations for unsupervised video prediction. In: IROS, pp. 5801–5806 (2018)
20. Jin, B., et al.: Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction. In: Proceedings of CVPR, pp. 4554–4563 (2020)
21. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Phys. Rev. E **69**(6), 066138 (2004)
22. Lee, A.X., et al.: Stochastic adversarial video prediction. arXiv preprint arXiv:1804.01523 (2018)

23. Lee, S., et al.: Video prediction recalling long-term motion context via memory alignment learning. In: Proceedings of CVPR, pp. 3054–3063 (2021)
24. Li, S., et al.: Semireward: a general reward model for semi-supervised learning. arXiv preprint arXiv:2310.03013 (2023)
25. Li, S., et al.: Moganet: multi-order gated aggregation network. In: The Twelfth International Conference on Learning Representations (2023)
26. Li, S., et al.: Masked modeling for self-supervised representation learning on vision and beyond. arXiv preprint arXiv:2401.00897 (2023)
27. Lin, T., Zha, H.: Riemannian manifold learning. IEEE Trans. Pattern Anal. Mach. Intell. **30**(5), 796–809 (2008)
28. Liu, X., Yin, J., Liu, J., Ding, P., Liu, J., Liu, H.: Trajectorycnn: a new spatio-temporal feature learning network for human motion prediction. IEEE Trans. Circuits Syst. Video Technol. **31**(6), 2133–2146 (2020)
29. Lotter, W., Kreiman, G., Cox, D.: Deep predictive coding networks for video prediction and unsupervised learning. In: ICLR (2017)
30. Mahmoud, A., Mohammed, A.: A survey on deep learning for time-series forecasting. In: Hassanien, A.E., Darwish, A. (eds.) Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges. SBD, vol. 77, pp. 365–392. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-59338-4_19
31. Marc'Aurelio Ranzato, A.S., Bruna, J., Mathieu, M., Collobert, R., Chopra, S.: Video (language) modeling: a baseline for generative models of natural videos. CoRR arxiv:1412.66042 (2014)
32. Mendelson, B.: Introduction to Topology. Courier Corporation, North Chelmsford (1990)
33. Oliu, M., Selva, J., Escalera, S.: Folded recurrent neural networks for future video prediction. In: Proceedings of ECCV, pp. 716–731 (2018)
34. Oprea, S., et al.: A review on deep learning techniques for video prediction. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
35. Patraucean, V., Handa, A., Cipolla, R.: Spatio-temporal video autoencoder with differentiable memory. arXiv preprint arXiv:1511.06309 (2015)
36. Schuldt, C., et al.: Recognizing human actions: a local SVM approach. In: ICPR, vol. 3, pp. 32–36 (2004)
37. Shi, X., et al.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. Adv. NIPS **28** (2015)
38. Simmons, G.F.: Introduction to topology and modern analysis, vol. 44. Tokyo (1963)
39. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using LSTMs. In: ICML, pp. 843–852 (2015)
40. Su, J., et al.: Convolutional tensor-train LSTM for spatio-temporal learning. Adv. NIPS **33**, 13714–13726 (2020)
41. Tan, C., Gao, Z., Li, S., Li, S.Z.: Simvp: towards simple yet powerful spatiotemporal predictive learning. arXiv preprint arXiv:2211.12509 (2022)
42. Tan, C., et al.: Openstl: a comprehensive benchmark of spatio-temporal predictive learning. Adv. Neural. Inf. Process. Syst. **36**, 69819–69831 (2023)
43. Tan, C., et al.: Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. arXiv preprint arXiv:2311.14109 (2023)
44. Tan, C., Xia, J., Wu, L., Li, S.Z.: Co-learning: learning from noisy labels with self-supervision. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 1405–1413 (2021)

45. Tan, C., et al.: Temporal attention unit: towards efficient spatiotemporal predictive learning. arXiv preprint arXiv:2206.12126 (2022)
46. Tschannen, M., Djolonga, J., Rubenstein, P.K., Gelly, S., Lucic, M.: On mutual information maximization for representation learning. In: International Conference on Learning Representations (2019)
47. Tulyakov, et al.: Mocogan: decomposing motion and content for video generation. In: Proceedings of CVPR, pp. 1526–1535 (2018)
48. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. 5998–6008 (2017)
49. Villegas, R., et al.: Decomposing motion and content for natural video sequence prediction. In: ICLR (2017)
50. Wang, Y., et al.: Predrnn: recurrent neural networks for predictive learning using spatiotemporal LSTMs. Adv. NIPS **30** (2017)
51. Wang, Y., et al.: Eidetic 3D LSTM: a model for video prediction and beyond. In: ICLR (2018)
52. Wang, Y., et al.: Predrnn++: towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In: ICML, pp. 5123–5132 (2018)
53. Wang, Y., et al.: Memory in memory: a predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In: Proceedings of CVPR, pp. 9154–9162 (2019)
54. Wei, J., et al.: Enhancing human-like multi-modal reasoning: a new challenging dataset and comprehensive framework. arXiv preprint arXiv:2307.12626 (2023)
55. Wu, H., et al.: Motionrnn: a flexible model for video prediction with spacetime-varying motions. In: Proceedings of CVPR, pp. 15435–15444 (2021)
56. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
57. Xu, Z., Wang, Y., Long, M., Wang, J., KLiss, M.: Predcnn: predictive learning with cascade convolutions. In: IJCAI, pp. 2940–2947 (2018)
58. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
59. Yu, W., et al.: Efficient and information-preserving future frame prediction and beyond. In: ICLR (2019)