State-space Decomposition Model for Video Prediction Considering Long-term Motion Trend

Fei Cui[1] , **Jiaojiao Fang**[1] , **Xiaojiang Wu**[1] , **Zelong Lai**[1] , **Mengke Yang**[1] , **Menghan Jia**[1] and **Guizhong Liu**[1*]

[1]Xi'an Jiaotong University

arXiv:2404.11576v1 [cs.CV] 17 Apr 2024

Abstract

Stochastic video prediction enables the consideration of uncertainty in future motion, thereby providing a better reflection of the dynamic nature of the environment. Stochastic video prediction methods based on image auto-regressive recurrent models need to feed their predictions back into the latent space. Conversely, the state-space models, which decouple frame synthesis and temporal prediction, proves to be more efficient. However, inferring long-term temporal information about motion and generalizing to dynamic scenarios under non-stationary assumptions remains an unresolved challenge. In this paper, we propose a state-space decomposition stochastic video prediction model that decomposes the overall video frame generation into deterministic appearance prediction and stochastic motion prediction. Through adaptive decomposition, the model's generalization capability to dynamic scenarios is enhanced. In the context of motion prediction, obtaining a prior on the long-term trend of future motion is crucial. Thus, in the stochastic motion prediction branch, we infer the long-term motion trend from conditional frames to guide the generation of future frames that exhibit high consistency with the conditional frames. Experimental results demonstrate that our model outperforms baselines on multiple datasets.

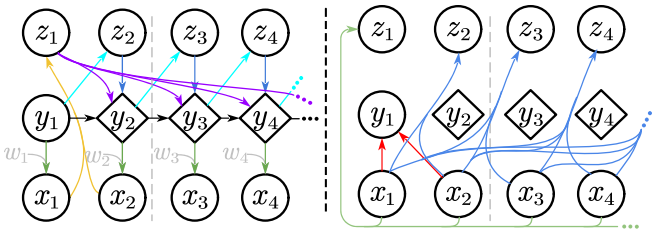
1 Introduction

Video prediction involves capturing the implicit environmental dynamics embedded in videos, aligning with the prior knowledge of model-based reinforcement learning. Therefore, reasonable predictions about the future from conditional frames have many applications in decision tasks [Finn and Levine, 2017; Piergiovanni *et al.*, 2019; Dugas *et al.*, 2022]. Video prediction aims to capture the dynamic representation of the world by modeling the prior knowledge of how the environment operates. Given the inherent stochastic nature of the world [Denton and Fergus, 2018], deterministic approaches for video prediction [Wang *et al.*, 2017; Jin *et al.*, 2020; Wu *et al.*, 2020; Gao *et al.*, 2019] fall short of capturing the complete dynamics of the environment. On the contrary, stochastic video prediction [Denton and Fergus, 2018; Franceschi *et al.*, 2020; Akan *et al.*, 2021], not relying on deterministic generation rules, exhibits superior generalization ability.

In stochastic video prediction, the key lies in how to capture the implicit motion cues embedded in the video. In contrast to the background (such as static room layout and indoor furniture) exhibiting shifts with camera movements, the complexity of motion subjects (such as pedestrians and moving cars) is higher and characterized by randomness. Traditional motion predictions often adopt deterministic approaches to forecast changes in motion, neglecting the various plausible possibilities of future motion. Alternatively, some predictions assume that the background is stationary, limiting the applicability of prediction models in fields such as navigation and autonomous driving. SLAMP [Akan *et al.*, 2021] decomposes stochastic video prediction into appearance-motion components but does not consider the long-term history of motion. Our insight is that the future development of motion has stochasticity, while backgrounds, such as static room layout, furniture, etc., exhibit deterministic shifts over time. Predicting the movement of dynamic subjects like pedestrians, who possess their independent consciousness, presents a challenging task. Therefore, we propose to decompose the overall video prediction into deterministic appearance prediction and stochastic motion prediction, aiming to adaptively focus on challenging-to-predict parts of the dynamics and sample future motion possibilities from the predicted distribution. Similar to SRVP [Franceschi *et al.*, 2020], our stochastic motion prediction branch relies on learning the residual updates of the latent states for stochastic variables to learn the system's temporal evolution. Deterministic appearance prediction, on the other hand, involves determining the background's temporal evolution based on deterministic residual updates.

In videos, there is implicit long-term historical information about motion. We aim to predict reasonable future frames based on the given conditional frames. Humans often make reasonable predictions about the future based on a few given frames because they match the temporally historical information inferred from the conditional frames with their long-term memory, allowing them to anticipate what will happen next. Therefore, inferring long-term motion trend from conditional

*Corresponding author



(a) generative model p (b) inference model q

Figure 1: Generative model p and Inference model q of our method, where circles and diamonds represent stochastic and deterministic variables, respectively. (a) In the generative model, the global motion trend variable $\boldsymbol{z}_1$ is generated from the conditional frames $\boldsymbol{x}_{1:k}$ (here $k = 2$), and the local dynamic variable $\boldsymbol{z}_t$ is generated from the previous motion variable $\boldsymbol{y}_{t-1}$. (b) In the inference model, $\boldsymbol{z}_1$ is inferred from the complete sequence $\boldsymbol{x}_{1:T}$, and the local dynamic $\boldsymbol{z}_t$ is inferred from the frame sequence $\boldsymbol{x}_{1:t}$. The motion variable $\boldsymbol{y}_t$ and the appearance variable $\boldsymbol{w}_t$ are jointly decoded to generate the frame $\hat{\boldsymbol{x}}_t$.

frames is crucial for predicting the future. To achieve this, we infer the overall long-term motion trend from the complete input sequence as the global dynamic to assist in predicting the inner-frame transition in stochastic motion prediction branch.

Our contributions are summarized as follows:

- We propose a state-space decomposition video prediction model that decomposes frame prediction into stochastic motion prediction and deterministic appearance prediction. Building upon a Gaussian prior for motion variables, the deterministic appearance prediction branch adaptively focuses on static features in frames.
- Our model utilizes a temporal transformer to infer the prior of global dynamics from the conditional frames, approximating the long-term motion trend. This results in the generation of future long-term sequence consistent with the ground truth.
- Experimental results demonstrate that the proposed approach achieves state-of-the-art performance on multiple datasets for the task of stochastic video prediction.

2 Related Work

2.1 Future Frame Video Prediction

Video prediction aims to predict future frames from a few conditional frames, requiring the extraction of reasonable motion clues from the image frames. The temporal relationship implicit in the sequence of image frames is a focal point in the prediction task. Previous works have predominantly modeled the temporal dynamics from raw pixel frames [Vondrick and Torralba, 2017; Chatterjee *et al.*, 2021; Ye and Bilodeau, 2023] or optical flow [Walker *et al.*, 2016; Liang *et al.*, 2017; Gao *et al.*, 2019; Akan *et al.*, 2021]. Recurrent neural networks [Wang *et al.*, 2018; Jin *et al.*, 2020] or transformers [Ye and Bilodeau, 2023; Farazi *et al.*, 2021] have been commonly employed to infer motion in prior works. We also employ a recurrent model to infer frame-to-frame changes. The generation of video prediction can be deterministic [Xu *et al.*, 2018; Chang *et al.*, 2022], but it struggles to capture the uncertainties present in the real world. Modeling the randomness of the future through latent variables and sampling from the predicted future distribution to generate upcoming image frames [Denton and Fergus, 2018; Franceschi *et al.*, 2020] aligns more closely with real-world physical priors. Therefore, our work also captures complex future motion trends by predicting the distribution of future motion variables. This encompasses both local motion trends (frame-to-frame dynamics) and long-term motion trend (overall motion types, such as running, dancing, etc.).

2.2 State-space Models

The state-space model excels in modeling temporal sequences in a low-dimensional latent space [Hafner *et al.*, 2019; Karl *et al.*, 2016; Gregor *et al.*, 2018; Franceschi *et al.*, 2020; Goel *et al.*, 2022; Newman *et al.*, 2023]. Unlike many auto-regressive models [Weissenborn *et al.*, 2019; Kalchbrenner *et al.*, 2017; Micheli *et al.*, 2022; Denton and Fergus, 2018; Akan *et al.*, 2021] commonly employed in prediction tasks, the temporal sequence in a state-space model flows in the latent space, decoupling the tasks of temporal prediction and frame generation. Consequently, the state-space model reduces the dependence on the encoder for generating the next frame. The state-of-the-art state-space model SRVP [Franceschi *et al.*, 2020] models future sequence by predicting frame-to-frame residual terms, generating dynamically continuous sequences of future images. However, its content variables are inferred solely from the initial frame, limiting its applicability to tasks with a static background. In our approach, we also adopt a state-space model but predict the appearance variables that evolve over time, enabling adaptation to complex and dynamic scenarios. SLAMP [Akan *et al.*, 2021] decouples motion and appearance prediction, explicitly modeling the local motion history (optical flow) to predict the next frame. However, it does not consider the long-term motion trend. In our stochastic motion prediction, we introduce global motion constraints aimed at guiding the generation of future long-term sequence consistent with long-term motion trend.

3 Method

3.1 Stochastic Motion Prediction

The image frame sequence contains the dynamic features of moving subjects (such as motion type and speed of subjects) as well as the static features of the background (such as streets, trees, etc.). We define motion variable $\boldsymbol{y}$ and appearance variable $\boldsymbol{w}$, representing dynamic features related to the moving subject and static features associated with the background in the image frames, respectively. Consequently, at time step t, the original pixel frame $\boldsymbol{x}_t$ can be decoded using the motion variable $\boldsymbol{y}_t$ and the appearance variable $\boldsymbol{w}_t$ jointly. The image frame sequence $\boldsymbol{x}_{1:T}$ is encoded to derive the motion variables $\boldsymbol{y}_{1:T}$. Following the setup of the state-space model, the video dynamics evolve over time in the latent space of the motion variables. At time step t, the next motion variable $\boldsymbol{y}_{t+1}$ explicitly depends on current motion variable $\boldsymbol{y}_t$. Adhering to physical priors, the the motion of

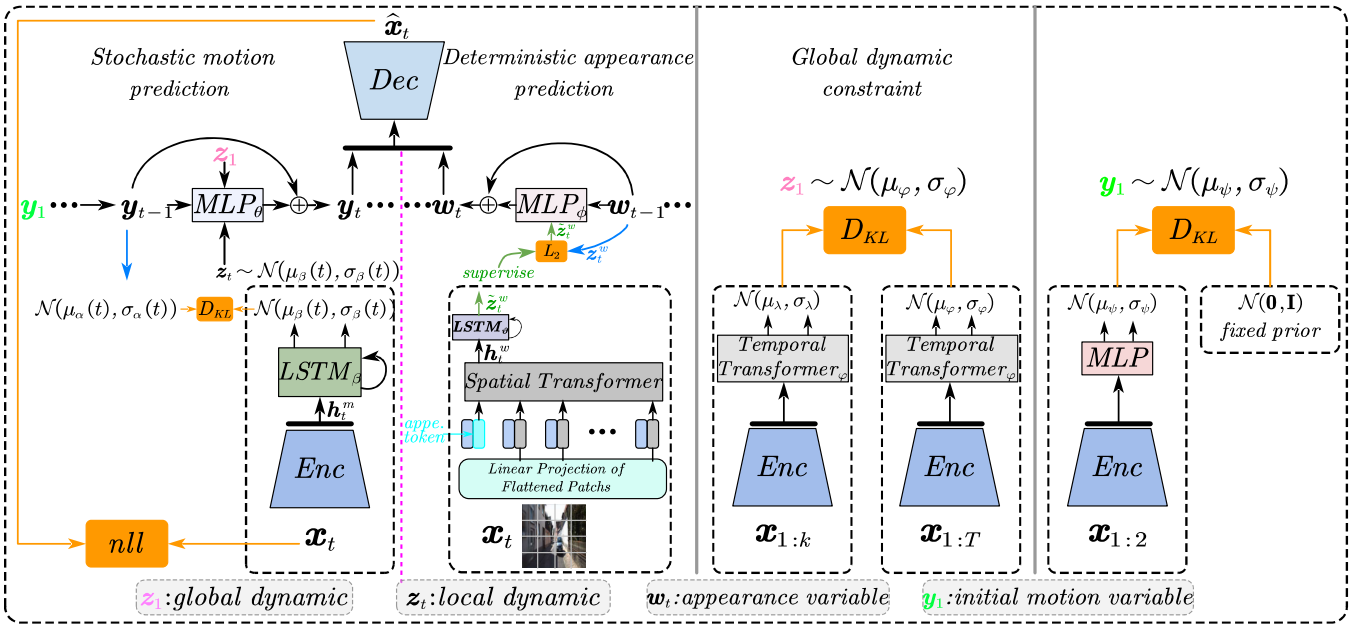


Figure 2: Framework of our method. The original frames $\boldsymbol{x}_{1:T}$ are mapped to a latent space through an encoder, and a LSTM captures the temporal dynamics within this latent space in the motion prediction branch. In the appearance prediction branch, a ViT is employed to encode the static features related to the background. To encourage the motion varibales to disregard static features, a standard Gaussian prior is applied to the motion variables (**right**). The prior and posterior of the global dynamic variable $\boldsymbol{z}_1$ are inferred from the conditional frames $\boldsymbol{x}_{1:k}$ and input frames $\boldsymbol{x}_{1:T}$, respectively (**middle**). The frame $\hat{\boldsymbol{x}}_t$ is jointly decoded from the appearance variable $\boldsymbol{w}_t$ and the motion variable $\boldsymbol{y}_t$ (**left**). The training pipeline and testing pipeline are detailed in Appendix B.

moving subject not only needs to adhere to the current local motion trend (i.e., inter frame local dynamic) but also must align with the long-term motion trend (i.e., global motion context). For instance, predicting long-term future frames under the condition of person running frames requires alignment with the person's long-term running context. Therefore, we define variable $\boldsymbol{z}_t$ $(2 \leq t \leq T)$ and $\boldsymbol{z}_1$ to represent the inter-frame local dynamic of frame x_t with respect to x_{t-1} and the global motion trend embedded in the video, respectively. This temporal evolution of video dynamics is modeled using a residual network (ie., MLP_θ) to encapsulate the frame-to-frame transition as follows:

$$\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \mathrm{MLP}_\theta(\boldsymbol{y}_t, \boldsymbol{z}_{t+1}, \boldsymbol{z}_1) \tag{1}$$

The overall generative model is illustrated in Figure 1(a). At time step t, The local motion trend $\boldsymbol{z}_{t+1}$ is generated by the motion variable $\boldsymbol{y}_t$, denoted as $\boldsymbol{z}_{t+1} \sim \mathcal{N}(\mu_\alpha(\boldsymbol{y}_t), \sigma_\alpha(\boldsymbol{y}_t))$. Simultaneously, the global $\boldsymbol{z}_1$ is generated from the entire sequence of conditional frames $\boldsymbol{x}_{1:k}$, indicated by $\boldsymbol{z}_1 \sim \mathcal{N}(\mu_\lambda(\boldsymbol{y}_{1:k}), \sigma_\lambda(\boldsymbol{y}_{1:k}))$, where k is the length of the conditional frames. To encourage motion variables to focus on dynamic features of the moving subject, we apply a standard Gaussian prior to the initial motion variable y_1 to discard unnecessary information, i.e, $\boldsymbol{y}_1 \sim \mathcal{N}(0, I)$. The pixel frame $\boldsymbol{x}_t$ are generated by both the motion variable $\boldsymbol{y}_t$ and the appearance variable $\boldsymbol{w}_t$, expressed as $\hat{\boldsymbol{x}}_t = Dec([\boldsymbol{w}_t, \boldsymbol{y}_t])$.

3.2 Deterministic Appearance Prediction

In video prediction, moving subjects such as pedestrians and vehicles exhibit intricate motion patterns, reflecting the uncertainties present in the real world. Conversely, static objects in video streams, like room layouts, streets, and trees, remain stationary and may appear shift due to camera motion. Some existing works [Denton and Birodkar, 2017; Franceschi *et al.*, 2020] separate the inference of content variables from motion prediction, but the assumption of invariant content variables constrains their performance in dynamic scenarios. Our insight is that static features associated with the background (such as the position, appearance, spatial relationships of static objects) undergo deterministic shifts over time, relying on the shift of the background. Hence, we define the variable $\boldsymbol{z}^w_{t+1}$ to represent inter-frame transition dynamic of appearance feature $\boldsymbol{w}_{t+1}$ with respect to $\boldsymbol{w}_t$. An MLP is used to predict appearance variables evolving over time.

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \mathrm{MLP}_\phi(\boldsymbol{w}_t, \boldsymbol{z}^w_{t+1}) \tag{2}$$

Unlike the prediction of stochastic distributions of motion variables, appearance variables transit deterministically. In other words, $\boldsymbol{z}^w_{t+1}$ is directly predicted from the previous appearance variable rather than sampled from a predicted distribution. Additionally, in contrast to the strong Gaussian prior for initial motion variable $\boldsymbol{y}_1$, the initial appearance variable $\boldsymbol{w}_1$ is directly encoded from the initial frame $\boldsymbol{x}_1$.

The results from SRVP [Franceschi *et al.*, 2020] have demonstrate that the strong Gaussian prior for initial motion variables encourages the motion variables to focus on the complex motion of the moving subjects, thereby neglecting unnecessary information for the motion of subjects. We employ a Vision Transformer (ViT) to encode appearance vari-

ables from pixel frames, as detailed in Section 3.4. The learnable appearance token encourages the ViT encoder to adaptively focus on some static background information in the pixel frames.

3.3 Variational Inference

After deriving the appearance variables $\boldsymbol{w}$, the complete evidence lower bound (ELBO) of the model can be derived. The conditional joint probability corresponding to the generative graph model shown in Figure 1(a) is given by:

$$\begin{aligned} &p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}, \boldsymbol{y}_{1:T} | \boldsymbol{w}_{1:T}) = p(\boldsymbol{z}_1 | \boldsymbol{x}_{1:k}) p(\boldsymbol{y}_1) \\ &\prod_{t=2}^{T} p(\boldsymbol{z}_t | \boldsymbol{y}_{t-1}) p(\boldsymbol{y}_t | \boldsymbol{y}_{t-1}, \boldsymbol{z}_t, \boldsymbol{z}_1) \prod_{t=1}^{T} p(\boldsymbol{x}_t | \boldsymbol{y}_t, \boldsymbol{w}_t) \end{aligned} \tag{3}$$

It can be observed that the conditional joint probability depends on the initial motion variable $\boldsymbol{y}_1$, the global dynamic variable $\boldsymbol{z}_1$, and the local dynamic variables $\boldsymbol{z}_{2:T}$.

The latent variable $\boldsymbol{z}_1$ should reflect the long-term motion trend embedded in video, such as motion type (running or walking), direction, etc. Therefore, $\boldsymbol{z}_1$ is inferred from the entire pixel frame sequence $\boldsymbol{x}_{1:T}$, and the latent variables $\boldsymbol{z}_t$ ($2 \leq t \leq T$) are encouraged to reflect inter-frame dynamics in the video sequence. The local motion trend depends on the current frame and previous frames, and similar to prior works [Franceschi *et al.*, 2020; Denton and Fergus, 2018; Akan *et al.*, 2021], $\boldsymbol{z}_t$ is inferred from the frame sequence $\boldsymbol{x}_{1:t}$, while the initial motion variable $\boldsymbol{y}_1$ are inferred from the initial two frames. To fit the true prior distribution of the latent variables as closely as possible, we employ a deep variational inference model to simulate the distributions of various latent variables. The overall inference model is designed as shown in Figure 1(b), and the variational distribution obtained from the inference model is:

$$\begin{aligned} q_{Z,Y} = q(\boldsymbol{z}_{1:T}, \boldsymbol{y}_{1:T} | \boldsymbol{x}_{1:T}, \boldsymbol{w}_{1:T}) = q(\boldsymbol{z}_1 | \boldsymbol{x}_{1:T}) q(\boldsymbol{y}_1 | \boldsymbol{x}_{1:2}) \\ \prod_{t=2}^{T} q(\boldsymbol{z}_t | \boldsymbol{x}_{1:t}) q(\boldsymbol{y}_t | \boldsymbol{y}_{t-1}, \boldsymbol{z}_t, \boldsymbol{z}_1) \end{aligned} \tag{4}$$

Combining the variational distributions from the inference model, the lower bound of the likelihood for the pixel frame sequence $\boldsymbol{x}_{1:T}$ can be obtained (complete derivation is provided in Appendix A):

$$\begin{aligned} &\log p(\boldsymbol{x}_{1:T} | \boldsymbol{w}) = \int_{\boldsymbol{z}} \int_{\boldsymbol{y}} q(\boldsymbol{z}_{1:T}, \boldsymbol{y}_{1:T} | \boldsymbol{x}_{1:T}, \boldsymbol{w}) \log p(\boldsymbol{x}_{1:T} | \boldsymbol{w}) \mathrm{d}_{\boldsymbol{z}} \mathrm{d}_{\boldsymbol{y}} \\ &= \mathbb{E}_{(\boldsymbol{z}_{1:T}, \boldsymbol{y}_{1:T}) \sim q_{Z,Y}} [\log p(\boldsymbol{x}_t | \boldsymbol{y}_t, \boldsymbol{w}_t)] - D_{KL}(q(\boldsymbol{y}_1 | \boldsymbol{x}_{1:2}) || p(\boldsymbol{y}_1)) \\ &\quad - \mathbb{E}_{(\boldsymbol{z}_{1:T}, \boldsymbol{y}_{1:T}) \sim q_{Z,Y}} \sum_{t=2}^{T} D_{KL}[q(\boldsymbol{z}_t | \boldsymbol{x}_{1:t}) || p(\boldsymbol{z}_t | \boldsymbol{y}_{t-1})] \\ &\quad - D_{KL}(q(\boldsymbol{z}_1 | \boldsymbol{x}_{1:T}) || p(\boldsymbol{z}_1 | \boldsymbol{x}_{1:k})) \end{aligned} \tag{5}$$

Here, D_{KL} represents the KL divergence [Kullback and Leibler, 1951]. For the initial motion variable, we adopt a strong Gaussian prior aiming to encourage the motion variables to discard unnecessary information. For optimizing the log-likelihood $\log p(\boldsymbol{x}_t | \boldsymbol{y}_t, \boldsymbol{w}_t)$, we compute the gradient by calculating the negative log density function of $\hat{\boldsymbol{x}}_t$ with respect to a normal distribution created by ground truth $\boldsymbol{x}_t$. For the remaining KL divergence terms, we compute the gradient using the reparameterization technique [Kingma and Welling, 2013]. For more training details, please refer to Appendix B.

3.4 Architecture

The overall framework of our method is illustrated in Figure 2. In the stochastic motion prediction branch, to infer the local dynamics $\boldsymbol{z}_t$, we initially employ a convolutional neural network to encode frames into $\boldsymbol{h}^m_{1:T}$, i.e., $\boldsymbol{h}^m_{1:T} = \mathrm{Enc}(\boldsymbol{x}_{1:T})$. Subsequently, we use a LSTM to infer the posterior of z_t in a feed-forward fashion:

$$\begin{aligned} \boldsymbol{g}_t &= \mathrm{LSTM}_\beta(\boldsymbol{h}^m_{1:t})) \\ \mu_\beta(t), \sigma_\beta(t) &= \mathrm{MLP}(\boldsymbol{g}_t) \end{aligned} \tag{6}$$

As described in Section 3.1, we derive the prior of local dynamic $\boldsymbol{z}_t$ from the previous motion variable, denoted as :

$$\mu_\alpha(t), \sigma_\alpha(t) = \mathrm{MLP}(\boldsymbol{y}_{t-1}) \tag{7}$$

For the initial motion variable $\boldsymbol{y}_1$, we infer it using the first two frames, denoted as $\mu_\psi, \sigma_\psi = \mathrm{MLP}(\boldsymbol{h}^m_{1:2})$. Regarding the global dynamic $\boldsymbol{z}_1$, our perspective is that overall motion trend, type must be inferred through the complete long-term sequence. Therefore, we infer the posterior of $\boldsymbol{z}_1$ using the complete sequence with a temporal transfomer. During testing, when the complete sequence is not visible, we use the conditional frames to generate the prior of $\boldsymbol{z}_1$:

$$\begin{aligned} \mu_\varphi, \sigma_\varphi &= \mathrm{Transformer}_\varphi(\boldsymbol{h}^m_{1:T}) \\ \mu_\lambda, \sigma_\lambda &= \mathrm{Transformer}_\varphi(\boldsymbol{h}^m_{1:k}) \end{aligned} \tag{8}$$

For the deterministic appearance prediction branch, previous works [Arnab *et al.*, 2021; Ye and Bilodeau, 2023] demonstrate the advantages of Vision Transformer (ViT) in adaptively extracting image features. We employ a ViT as the appearance encoder to encode features related to the background. Building upon the strong Gaussian prior for motion variables to discard unnecessary information for motion, the learnable appearance token encourages ViT to adaptively focus on pixel patches related to the background, i.e., $\boldsymbol{h}^w_t = \mathrm{ViT}(\boldsymbol{x}_t)$. where $\boldsymbol{h}^w_t$ is the static features encoded by ViT, then a LSTM is used to generate inter-frame transition dynamics of appearance features, i.e., $\tilde{\boldsymbol{z}}^w_t = \mathrm{LSTM}_\vartheta(\boldsymbol{h}^w_{1:t})$. During testing, $\boldsymbol{z}^w_t$ is predicted from the previous appearance variable. During training, $\tilde{\boldsymbol{z}}^w_t$ is used to supervise $\boldsymbol{z}^w_t$ through $\mathcal{L}_2$ loss $\sum_{t=2}^{T} \|\tilde{\boldsymbol{z}}^w_t - \boldsymbol{z}^w_t\|_2$. Additionally, to facilitate $\boldsymbol{z}_{2:T}$ capturing local inter-frame dynamics, FlowDec with the same architecture as Dec is employed to decode optical flow $\boldsymbol{f}_{2:T}$ from the output $\boldsymbol{g}_{2:T}$ of the LSTM_β and warp it into frame $\tilde{\boldsymbol{x}}_{2:T}$ via differentiable warping [Jaderberg *et al.*, 2015].

$$\begin{aligned} \boldsymbol{f}(t) &= \mathrm{FlowDec}(\boldsymbol{g}_t) \\ \tilde{\boldsymbol{x}}_t &= \mathrm{warp}(\boldsymbol{f}(t), \boldsymbol{x}_{t-1}) \end{aligned} \tag{9}$$

where FlowDec is trained using $\mathcal{L}_2$ loss $\sum_{t=2}^{T} \|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}_t\|_2$. Please note that predicting optical flow is solely intended to encourage $\boldsymbol{z}_t$ to focus on the inter-frame local dynamic of frame $\boldsymbol{x}_t$ with respect to $\boldsymbol{x}_{t-1}$ during training. The model does not require optical flow to generate future frames.

4 Experiments

4.1 Datasets

To evaluate the proposed method's effectiveness across diverse scenarios, experiments are conducted on datasets with

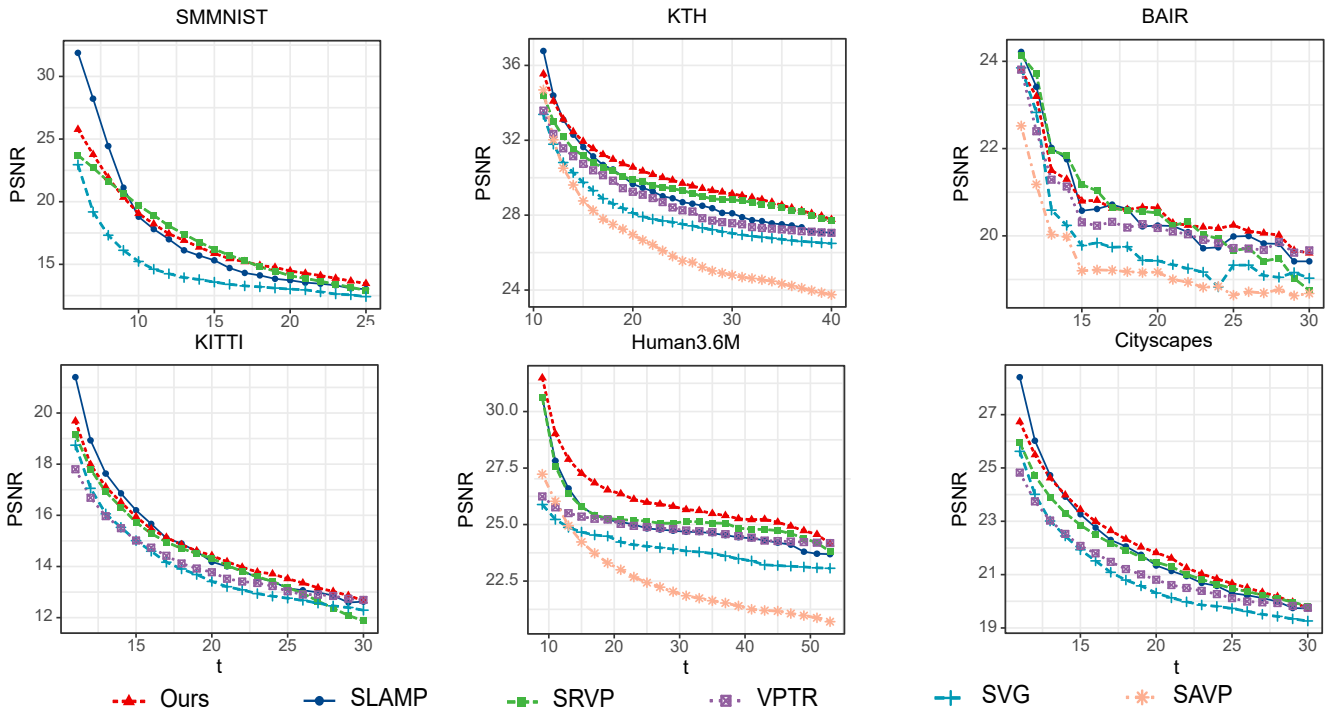


Figure 3: The PSNR scores over timestep t for our proposed method and various baselines. Each score represents the mean value obtained from five different samples generated by the models. Our proposed model achieved superior performance on the KTH, Human3.6M and Cityscapes datasets, while demonstrating comparable performance to state-of-the-art models on the BAIR, SMMNIST and KITTI datasets in terms of the PSNR metric.

both stationary and dynamic backgrounds, including SMMNIST [Denton and Fergus, 2018], BAIR [Ebert *et al.*, 2017], KTH [Schuldt *et al.*, 2004], Human3.6M [Ionescu *et al.*, 2011], Cityscapes [Cordts *et al.*, 2016], and KITTI [Geiger *et al.*, 2012]. SMMNIST involves moving two MNIST digits [LeCun *et al.*, 1998] with possible overlap. The BAIR Push dataset features a robotic arm pushing an object to induce movement. The KTH dataset encompasses human motion videos with various patterns like walking and running. The Human3.6M dataset includes videos of subjects performing actions indoors, exhibiting more variability than KTH. For assessing performance in complex real-world driving scenarios with moving backgrounds, we also evaluate our method on KITTI and Cityscapes datasets. KITTI footage was captured while driving in Karlsruhe, and Cityscapes includes street driving videos from multiple cities, offering more diverse driving environments than KITTI. More details about the datasets are provided in Appendix C.

4.2 Implementation Details

For the BAIR, KTH, SMMNIST, and Human3.6M datasets, image frames are 64×64 pixels, and for KITTI and Cityscapes datasets, frames are 128×128 pixels. During training, BAIR, KITTI, Cityscapes, and KTH use the initial 10 frames as conditionals for predicting the next 10 frames; Human3.6M uses the initial 8 frames to predict the subsequent 8 frames; SMMNIST uses the initial 5 frames to predict the next 10 frames. During testing, BAIR, Cityscapes, KITTI, and SMMNIST predict the next 20 frames. For KTH, 30 frames are predicted, and for Human3.6M, 45 frames are predicted. Our encoder-decoder adopts the VGG16 [Simonyan and Zisserman, 2014] architecture for a fair comparison with the previous methods [Franceschi *et al.*, 2020; Akan *et al.*, 2021]. Additional training details are available in Appendix B.

Baselines and Evaluation Metrics: To evaluate our state-space decomposition model for stochastic video prediction, we compare it with leading variational methods (SVG [Denton and Fergus, 2018], SAVP [Lee *et al.*, 2018], SRVP [Franceschi *et al.*, 2020], SLAMP [Akan *et al.*, 2021]) and a deterministic approach (VPTR [Ye and Bilodeau, 2023]) on various datasets. We use three standard metrics for future frame prediction performance: Peak Signal-to-Noise Ratio (PSNR) for reconstruction quality, Structural Similarity Index (SSIM) for structural alignment, and Learned Perceptual Image Patch Similarity (LPIPS [Zhang *et al.*, 2018]) for dissimilarity assessment based on feature maps.

4.3 Evaluation

Table 1 presents quantitative results of our method and baselines across various datasets. Our method achieves state-of-the-art (SOTA) performance across multiple datasets, particularly excelling in the SSIM metric. In KTH dataset, our method demonstrates superior performance in inferring the long-term motion context coherently with the conditional frames, as illustrated in Figure 4, in the case of human walking samples, our method accepts constraints from global dy-

Model	dataset:**SMMNIST**(5→20)			dataset:**BAIR**(10→20)		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
SVG	14.50±0.04	0.7090±0.0015	-	19.85±0.26	0.8301±0.0088	0.0553±0.0034
SAVP	-	-	-	19.39±0.25	0.8137±0.0092	0.0564±0.0026
VPTR	-	-	-	20.41±0.27	0.8363±0.0088	0.0560±0.0036
SRVP	16.93±0.07	0.7799±0.0020	-	20.63±0.28	0.8448±0.0078	0.0521±0.0032
SLAMP	**18.07±0.08**	0.7736±0.0019	-	**20.71±0.26**	0.8402±0.0086	0.0575±0.0037
Ours	17.12±0.07	**0.7809±0.0020**	-	20.67±0.28	**0.8459±0.0078**	**0.0520±0.0032**
Model	dataset:**KTH**(10→30)			dataset:**Human3.6M**(8→45)		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
SVG	28.06±0.29	0.8438±0.0054	0.0923±0.0038	23.94±0.19	0.8889±0.0028	0.0636±0.0018
SAVP	26.51±0.29	0.7564±0.0062	0.1120±0.0039	22.61±0.18	0.8036±0.0031	0.0764±0.0019
VPTR	28.77±0.31	0.8674±0.0055	0.0851±0.0035	24.82±0.21	0.8948±0.0026	0.0621±0.0018
SRVP	29.69±0.32	0.8697±0.0046	**0.0736±0.0029**	25.30±0.19	0.9074±0.0022	0.0509±0.0013
SLAMP	29.39±0.30	0.8646±0.0050	0.0795±0.0034	25.17±0.19	0.9032±0.0022	0.0549±0.0015
Ours	**30.30±0.31**	**0.8766±0.0045**	0.0743±0.0029	**26.07±0.20**	**0.9160±0.0021**	**0.0501±0.0013**
Model	dataset:**KITTI**(10→20)			dataset:**Cityscapes**(10→20)		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
SVG	13.97±0.47	0.3572±0.0183	0.5537±0.0379	20.94±0.61	0.6211±0.0218	0.3094±0.0209
VPTR	14.13±0.44	0.3558±0.0198	0.5438±0.0243	21.24±0.53	0.6279±0.0221	0.3214±0.0235
SRVP	14.53±0.34	0.3637±0.0195	0.5264±0.0235	21.77±0.44	0.6349±0.0161	0.3147±0.0145
SLAMP	**14.87±0.49**	0.3698±0.0207	0.4912±0.0397	22.01±0.71	0.6513±0.0232	**0.2937±0.0214**
Ours	14.67±0.46	**0.3781±0.0230**	**0.4572±0.0236**	**22.12±0.46**	**0.6555±0.0163**	0.3014±0.0134

Table 1: Numerical results (mean and 95%-confidence interval) for PSNR, SSIM, and LPIPS of our proposed method and baselines.

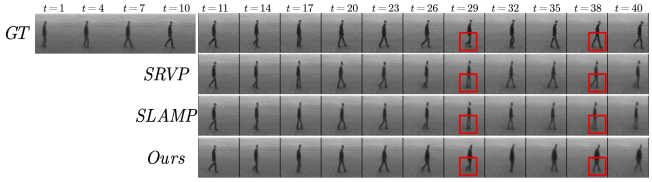


Figure 4: Person walking. The top row shows the ground truth, followed by the predictions from SRVP, SLAMP and our method.

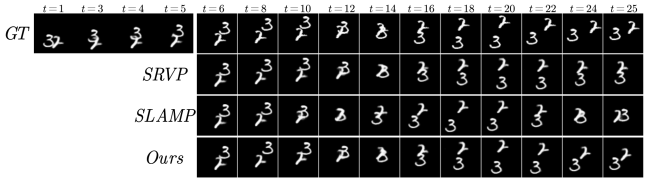


Figure 5: Overlapping digits. This figure shows two overlapping digits and the predictions from SRVP, SLAMP and our method.

namics, leading to more realistic leg details in the predictions. Specifically, on the Human3.6M dataset, where only 8 frames are used to predict the next 45 frames, our method outperforms baselines on all three metrics. The KITTI and Cityscapes datasets contain real driving videos with changing backgrounds over time. As indicated in Table 1, our method outperforms SRVP and SVG significantly on KITTI and Cityscapes datasets, achieving a level comparable to SLAMP while maintaining a computational advantage (see Table 2). For SMMNIST dataset, the challenge lies in the separation prediction of intersecting digits. As shown in Figure 5, our method benefits from the guidance of the global motion trend z_1, successfully decoupling two digits even after their intersection, and accurately predicting the motion direction of each digit. On the BAIR dataset, our method surpasses baseline methods in both PSNR and SSIM metrics, though it falls behind SRVP in terms of LPIPS. When predicting the long-term future, as shown in Figure 6, our method accurately predicts the long-term displacement of the robotic arm.

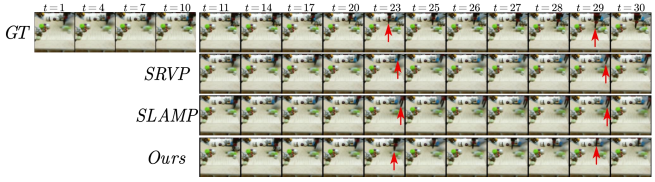


Figure 6: Results on BAIR. Our method accurately predicts the long-term displacement of the robotic arm compared to baselines.

In order to contrast the performance of our method and baselines at each time step, we plot the relationship between prediction quality and time steps in Figure 3. It cant be seen that on the KTH and Human3.6M datasets, the proposed method consistently outperforms baseline methods at each step. On the challenging KITTI and Cityscapes datasets, our method initially lags behind the state-of-the-art method SLAMP in the first few steps. However, it gradually surpasses SLAMP in later steps, owing to the decoupling of frame synthesis and temporal prediction by the state space model, resulting in smoother sequence predictions. Specific

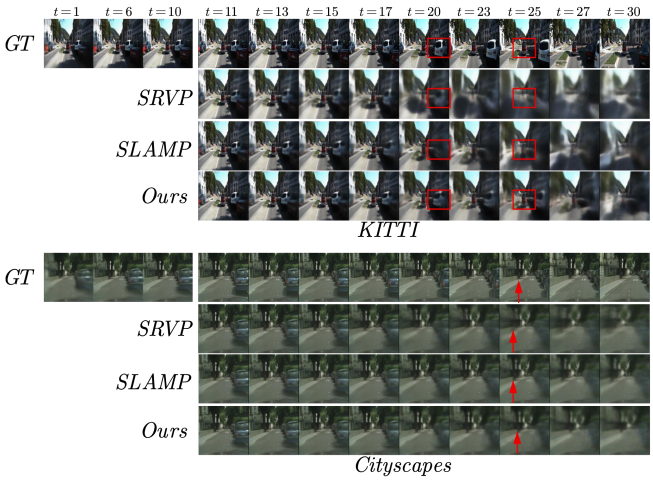


Figure 7: Results on KITTI (**up**) and Cityscapes (**down**). Our method predicts richer details (red square) and achieves more accurate positioning of the street trees (red arrow).

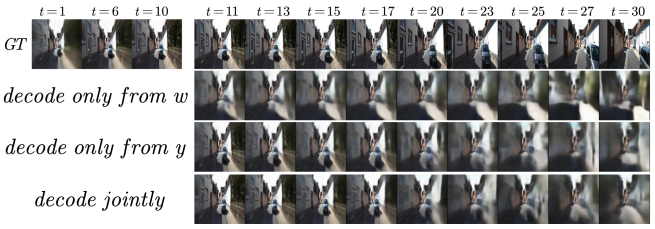


Figure 8: Results decoded separately. This figure shows videos decoded separately from w and y, and the result of joint decoding.

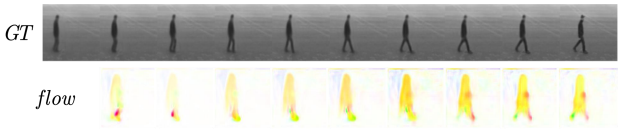


Figure 9: Oplical flow. This figure shows the pixel sequence and the opliacal flow decoded from the output of LSTM_β.

sequence visualizations are illustrated in Figure 7. For SM-MIST, SLAMP exhibits superior performance in the initial steps but experiences a rapid decline, falling short of SRVP and our method. In the BAIR dataset, our method achieves superior performance relative to baselines after step 20. Featuring a rapidly moving robotic arm with continuously changing directions, all the models have noticeable discontinuities.

4.4 Experiments on State-space Decomposition

Our method decomposes video prediction into stochastic motion prediction and deterministic appearance prediction. The Gaussian prior on the initial motion variable facilitates the attention of motion variables on the dynamic features of the subject. The learnable appearance token encourages the ViT to focus on static features related to the background. To validate the performance of different branches, we visualize sequences decoded separately from the appearance variable $\boldsymbol{w}$ and the motion variable $\boldsymbol{y}$ in Figure 8. It can be observed that the motion variables capture the movement of the subject (i.e., the car), including the motion direction and displacement, while appearance variables focuses more on static features such as background contours and spatial relationships. To further verify the ability of the stochastic motion prediction branch to model inter-frame local dynamics, we visualize the optical flow decoded from the output of LSTM_β in Figure 9. LSTM_β effectively captures local dynamics, demonstrating the rationality of predicting the distribution of local motion trend $z_{2:T}$ based on the output of LSTM_β in equation 6. Utilizing the ViT with a learnable appearance token allows capturing static information in the frame sequence. We also compare the case where VGG16 serves as the appearance encoder in Appendix D.

4.5 Ablation Studies

To further validate the impact of the appearance prediction branch and the global dynamics, we compare the performance in the following settings: (**i**) Ours (w/o $\boldsymbol{w}$), abandoning the appearance prediction branch and adopting the same static content variable scheme as SRVP [Franceschi *et al.*, 2020], (**ii**) Ours (w/o $\boldsymbol{z}_1$), Not considering the global dynamic $\boldsymbol{z}_1$ when predicting frame-to-frame transitions in the motion prediction branch, and (**iii**) Ours method, as described in Chapter 3. The results are presented in Table 2, where each component contributes to the predictive performance. These results indicate the effectiveness of the deterministic appearance prediction branch in adaptively encoding images and the global dynamic $\boldsymbol{z}_1$ for predicting long-term frame transitions. Additionally, the computational time increase introduced by the added components is marginal compared to SRVP. For more ablation experiments and visualization samples, please refer to Appendix D.

Method	PSNR↑	SSIM↑	LPIPS↓	Inf Time(s)↓
SRVP	14.53	0.3637	0.5264	**0.035**
SLAMP	**14.87**	0.3698	0.4912	0.369
Ours (w/o w)	<u>14.79</u>	0.3655	0.5068	<u>0.043</u>
Ours (w/o z_1)	14.50	<u>0.3723</u>	<u>0.4731</u>	0.049
Ours	14.67	**0.3781**	**0.4572**	0.058

Table 2: Ablation results on KITTI regarding PSNR, SSIM, LPIPS, and Inference Time (average inference time for testing 100 samples on RTX 2080Ti).

5 Conclusion

In this paper, we propose a state space decomposition video prediction model that decomposes the overall frame prediction into stochastic motion prediction and deterministic appearance prediction. The stochastic motion prediction module, when predicting inter-frame residuals, incorporates the global dynamics extracted from the conditional sequence to guide the motion predictions. Experimental results demonstrate that the proposed method achieves state-of-the-art performance on various datasets.

References

[Akan *et al.*, 2021] Adil Kaan Akan, Erkut Erdem, Aykut Erdem, and Fatma Güney. Slamp: Stochastic latent appearance and motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14728–14737, 2021.

[Arnab *et al.*, 2021] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 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.

[Chatterjee *et al.*, 2021] Moitreya Chatterjee, Narendra Ahuja, and Anoop Cherian. A hierarchical variational neural uncertainty model for stochastic video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9751–9761, 2021.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[Denton and Birodkar, 2017] Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. *Advances in Neural Information Processing Systems*, 30:1–10, 2017.

[Denton and Fergus, 2018] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International Conference on Machine Learning*, pages 1174–1183, 2018.

[Dugas *et al.*, 2022] Daniel Dugas, Olov Andersson, Roland Siegwart, and Jen Jen Chung. Navdreams: Towards camera-only rl navigation among humans. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2504–2511, 2022.

[Ebert *et al.*, 2017] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. *CoRL*, 12:16, 2017.

[Farazi *et al.*, 2021] Hafez Farazi, Jan Nogga, and Sven Behnke. Local frequency domain transformer networks for video prediction. In *2021 International Joint Conference on Neural Networks*, pages 1–10, 2021.

[Finn and Levine, 2017] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation*, pages 2786–2793, 2017.

[Franceschi *et al.*, 2020] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *International Conference on Machine Learning*, pages 3233–3246, 2020.

[Gao *et al.*, 2019] Hang Gao, Huazhe Xu, Qi-Zhi Cai, Ruth Wang, Fisher Yu, and Trevor Darrell. Disentangling propagation and generation for video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9006–9015, 2019.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[Goel *et al.*, 2022] Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's raw! audio generation with state-space models. In *International Conference on Machine Learning*, pages 7616–7633, 2022.

[Gregor *et al.*, 2018] Karol Gregor, George Papamakarios, Frederic Besse, Lars Buesing, and Theophane Weber. Temporal difference variational auto-encoder. *arXiv Preprint arXiv:1806.03107*, 2018.

[Hafner *et al.*, 2019] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565, 2019.

[Ionescu *et al.*, 2011] Catalin Ionescu, Fuxin Li, and Cristian Sminchisescu. Latent structured models for human pose estimation. In *2011 International Conference on Computer Vision*, pages 2220–2227, 2011.

[Jaderberg *et al.*, 2015] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[Jin *et al.*, 2020] Beibei Jin, Yu Hu, Qiankun Tang, Jingyu Niu, Zhiping Shi, Yinhe Han, and Xiaowei Li. Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4554–4563, 2020.

[Kalchbrenner *et al.*, 2017] Nal Kalchbrenner, Aäron Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *International Conference on Machine Learning*, pages 1771–1779, 2017.

[Karl *et al.*, 2016] Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick Van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. *arXiv Preprint arXiv:1605.06432*, 2016.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv Preprint arXiv:1312.6114*, 2013.

[Kullback and Leibler, 1951] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

[Liang *et al.*, 2017] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *proceedings of the IEEE International Conference on Computer Vision*, pages 1744–1752, 2017.

[Micheli *et al.*, 2022] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *The Eleventh International Conference on Learning Representations*, pages 1–21, 2022.

[Newman *et al.*, 2023] Ken Newman, Ruth King, Víctor Elvira, Perry de Valpine, Rachel S McCrea, and Byron JT Morgan. State-space models for ecological time-series data: Practical model-fitting. *Methods in Ecology and Evolution*, 14(1):26–42, 2023.

[Piergiovanni *et al.*, 2019] AJ Piergiovanni, Alan Wu, and Michael S Ryoo. Learning real-world robot policies by dreaming. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 7680–7687, 2019.

[Schuldt *et al.*, 2004] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 3, pages 32–36, 2004.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv Preprint arXiv:1409.1556*, 2014.

[Vondrick and Torralba, 2017] Carl Vondrick and Antonio Torralba. Generating the future with adversarial transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1020–1028, 2017.

[Walker *et al.*, 2016] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 835–851, 2016.

[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*, 30:1–10, 2017.

[Wang *et al.*, 2018] 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*, pages 1–14, 2018.

[Weissenborn *et al.*, 2019] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *International Conference on Learning Representations*, pages 1–24, 2019.

[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.

[Xu *et al.*, 2018] Jingwei Xu, Bingbing Ni, Zefan Li, Shuo Cheng, and Xiaokang Yang. Structure preserving video prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1460–1469, 2018.

[Ye and Bilodeau, 2023] Xi Ye and Guillaume-Alexandre Bilodeau. Video prediction by efficient transformers. *Image and Vision Computing*, 130:104612, 2023.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.