PredLDM: Spatiotemporal Sequence Prediction with Latent Diffusion Models

Anonymous authors
Paper under double-blind review

Abstract

Predicting the accurate and realistic future is an attractive landmark in spatiotemporal sequence prediction. Despite recent progress in spatiotemporal predictive models, explorations in this field are challenging due to difficulties in intricate global coherence and comprehensive history understanding. In this study, we introduce latent diffusion models (LDMs) into spatiotemporal sequence prediction (PredLDM) with a two-stage training paradigm. (i) To compress intricate global coherent spatiotemporal content into latent space, we propose the masked-attention transformer-based variational autoencoder (MT-VAE) by exploiting transformers with masked self-attention layers. (ii) Different from LDMs in generation-related fields where the condition in our problem settings is historical observations instead of texts, the condition-aware LDM (CA-LDM) is provided for comprehensive understanding of historical sequences. Our denoising diffusion process learns the distribution of both conditional generation and condition-aware reconstruction. Results on KittiCaltech, KTH and SEVIR datasets show that our PredLDM provides promising performance and realistic predictions in multiple scenarios including car driving, humans and weather evolutions. Code will be released here during camera ready.

1 Introduction

Spatiotemporal sequence prediction is a fundamental task in computer vision that given a sequence of images, neural networks predict the subsequent image sequence to describe what will happen in the future (Oprea et al., 2020; Shi et al., 2015). Different from video generation (Ho et al., 2022b) predicting from text prompts or unconditionally, this task is conditioned on historical observations of dynamic scenes (Oprea et al., 2020). By learning underlying spatiotemporal patterns from successive data with unsupervised manners, an ideal model is to predict accurate dynamics with realistic visual appearance (Lee et al., 2018). It can serve various disciplines, such as autonomous driving (Kwon & Park, 2019), robotics planning (Finn et al., 2016), traffic management (Liu et al., 2024) and weather forecasting (Zhang et al., 2023b).

For producing future frames, classical predictive models are mostly optimized by minimizing mean error between predictions and ground truth across spatial and temporal dimensions (Oprea et al., 2020). Shi et al. (2015) introduce ConvLSTM networks, which is a milestone at grasping spatiotemporal aspects with convolutional recurrent architectures. Inspired by this, advanced recurrent models (Wang et al., 2017; 2018; Wu et al., 2021b; Sun et al., 2023; Villegas et al., 2017; Oliu et al., 2018) and recurrent-free ones (Gao et al., 2022; Tan et al., 2023) emerge out. However, the mean error-related objective leads to generating blur for uncertain future outcomes (Oprea et al., 2020; Lee et al., 2018). To improve visual quality, although generative models like variational autoencoders (VAEs) (Villegas et al., 2019; Wu et al., 2021a; Babaeizadeh et al., 2021), generative adversarial networks (GANs) (Clark et al., 2019; Tulyakov et al., 2018) and flow-based models (Dorkenwald et al., 2021) are alternatives, they are easy for mode collapase and the performance is not satisfactory. As LDMs reveal promising performance with high-fidelity appearance especially in T2I (Nichol et al., 2021; Rombach et al., 2022) and T2V (Ho et al., 2022b; He et al., 2022) through learning joint distributions with conditions in latent space by iterative denoising diffusion processes (Rombach et al., 2022), we introduce LDMs into spatiotemporal sequence prediction, under consideration of intricate global coherence and comprehensive history understanding.

In the light of intricate global coherence, existing predictive models are restricted by finite-scale temporal variations within training samples, whereas temporal transformations are complex and nearly infinite in nature. For simulating intricate temporal patterns, a solution is to model as many diverse variations as possible in the pretraining stage (Devlin, 2018; He et al., 2022) by masked modeling (Xie et al., 2022; Cheng et al., 2022). This can serve perceptual compression in LDMs (Singer et al., 2022). Meanwhile, existing LDMs exploit 3D convolutions (Ho et al., 2022b; He et al., 2022) and convolutional temporal layers (Singer et al., 2022) to extend T2I models to T2V applications. This leads to global dependencies being neglected, limited by compression of the local receptive field of convolutions (Li et al., 2023). For modeling global reliance, transformers are natural alternatives. To solve intricate global coherence, we expect to propose a transformer-based VAE with masked modeling. In another light of comprehensive history understanding, different from generation-purpose models conditioned by text prompts, the condition in this problem setting is historical image sequences. Compared to text prompts describing scenes with highly dimensional symbols, conditions of spatiotemporal sequences are more difficult for machines to understand as raw pixels are low-level and diverse. It is expected to leave conditions comprehensively understood in latent space during denoising diffusion processes.

With respect to these problems, we propose a spatiotemporal predictive model called PredLDM. (i) To compress intricate global coherent spatiotemporal content into latent space, we propose exploiting transformer-based VAE with masked attention to capture complex and global coherence in MT-VAE. (ii) To comprehensively understand the historical observations, condition-aware latent diffusion is performed. The denoising diffusion process of CA-LDM learns the distribution for both conditional generation and condition-aware reconstruction.

Extensive experiments are conducted on KittiCaltech (Geiger et al., 2013), KTH (Schuldt et al., 2004) and SEVIR (Veillette et al., 2020) datasets. Results show accurate performance and realistic visual appearance of trained PredLDM, indicating the promising future of this study. Contributions can be summarized as:

- To predict the accurate and realistic future image sequences, we propose a spatiotemporal predictive model called PredLDM, by introducing LMDs into this field under consideration of intricate global coherent modeling and comprehensive history understanding.
- For intricate global coherence, MT-VAE is proposed by transformers with masked attention variationally compressing complex temporal patterns and global reliance.
- Considering comprehensive history understanding, CA-LDM is performed by learning distributions of both conditional generation and condition-aware reconstruction.
- Experiments on several datasets including KittiCaltech, KTH and SEVIR show superior performance of PredLDM with realistic appearance, revealing potential for continuous research and applications.

2 Related works

2.1 Spatiotemporal Sequence Prediction

Spatiotemporal sequence prediction produces the future sequence of images given by historical observations to describe what is going to happen (Oprea et al., 2020). This research direction originates from predictive coding (Huang & Rao, 2011; Rao & Ballard, 1999) which reveals the human behavior predicting visual signals through both space and time dimensions. Initial attempts from Ranzato et al. (2014) and Srivastava et al. (2015) introduce recurrent language baselines to model natural spatiotemporal signals. For explicitly modeling spatial information, Shi et al. (2015) propose using convolutions to replace fully connected layers in recurrent units. This attempt greatly inspires the progress on the recurrent predictive architectures (Wang et al., 2017; 2018; Wu et al., 2021b; Sun et al., 2023; Villegas et al., 2017; Oliu et al., 2018) on this task. Besides recurrently modeling, the sequence-to-sequence fashion (Gao et al., 2022; Tan et al., 2023) is employed with efficient U-Net structures to predict with a simplified configuration of convolutions. When minimizing mean error of predictions and uncertain future outcomes, these models usually generate blur appearance (Oprea et al., 2020). A straightforward solution is to exploit probabilistic models, like VAEs (Villegas et al., 2019;

Wu et al., 2021a; Babaeizadeh et al., 2021), GANs (Clark et al., 2019; Tulyakov et al., 2018) and flow-based ones (Dorkenwald et al., 2021). However, they are highly likely to lead mode collapse and hard to fit (Oprea et al., 2020). As LDMs are dominant in generation-related works (Rombach et al., 2022; Ho et al., 2022b), we introduce LDMs in spatiotemporal sequence prediction for potential explorations in this field.

2.2 Latent Diffusion Models

Diffusion models (DMs) are one of likelihood-based generative models, revealing first remarkable results in image generation communities (Ho et al., 2020; Nichol et al., 2021) by progressively reversing a Markov chain which iteratively adds noise to target distributions (Ronneberger et al., 2015). Benefiting from lower computational requirement and better expressivity than DMs (Song et al., 2020; Karras et al., 2022), recently LDMs (He et al., 2022; Rombach et al., 2022) have made tremendous breakthroughs on various tasks, including T2I generation (Nichol et al., 2021; Rombach et al., 2022; Balaji et al., 2022; Saharia et al., 2022; OpenAI, 2023; Midjourney, 2023; Peebles & Xie, 2023; Podell et al., 2023), T2V generation (Shi et al., 2015; He et al., 2022; Yan et al., 2021; Singer et al., 2022; Voleti et al., 2022; Ho et al., 2022a; Blattmann et al., 2023b; Zhou et al., 2022; Wang et al., 2023; Midjourney, 2023), text-to-audio generation (Liu et al., 2023), 3D shape generation (Vahdat et al., 2022), video editing (Liew et al., 2023), tabular data generation (Zhang et al., 2023a), video frame interpolation (Danier et al., 2024), etc. Most related directions are T2I and T2V models. In T2I generation (Nichol et al., 2021; Rombach et al., 2022; Balaji et al., 2022), novel images are generated with textual descriptions given as conditions, where representatives contain Dalle-2 (OpenAI, 2023), Midjourney (Midjourney, 2023), DiT (Peebles & Xie, 2023) and Stable Diffusion (Podell et al., 2023). T2V models are mostly inspired from T2I (Singer et al., 2022). VDM (Shi et al., 2015) reports first results by modifying 2D U-Net to a factorized 3D network to achieve video synthesis. Recent works include Imagen Video (Ho et al., 2022a), SORA (OpenAI, 2024), Make-A-Video (Singer et al., 2022), VideoGPT (Yan et al., 2021), MagicVideo (Zhou et al., 2022), Latte (Ma et al., 2024), StoryDiffusion (Zhou et al., 2024) and CogVideoX (Yang et al., 2024), extending existing image-based models to the video domain. They are usually built with two stages (Rombach et al., 2022). In the first stage, VQ-VAEs (Van Den Oord et al., 2017; Razavi et al., 2019) or VQ-GANs (Esser et al., 2021) are used as autoencoders to learn an expressive prior over discretized latent space. In the second stage, a denoising network commonly implemented by U-Net (Ho et al., 2020; Ronneberger et al., 2015; Dhariwal & Nichol, 2021) is trained to predict the less noisy video samples progressively, reversing the diffusion process where Gaussian noise is iteratively added onto the raw data with predefined timesteps. Different form T2I and T2V models, the condition of our PredLDM is historical observations. This difference makes the condition is not as easy as text prompts used to be, as the spatiotemporal content is low dimensional and highly complex.

2.3 Compression of Spatiotemporal Sequences

In order to compress spatiotemporal data, 3D convolutions are straightforward solutions (Ho et al., 2022b; Tran et al., 2015). Given multiple frames, VDM (Ho et al., 2022b) and LVDM (He et al., 2022) exploit 3D U-Net convolutions by replacing each 2D convolutions in image models with space-only 3D convolutions. Instead of 3D convolutions expensive at fitting or hard to train, 1D convolutional temporal layers are attractive combined with 2D convolutions (Tran et al., 2018). Make-A-Video (Singer et al., 2022) initializes the spatial convolutional layers with pretrained T2I weights and adds temporal convolutions to correlate spatial features across time dimensions, similar as Imagen Video (Ho et al., 2022a), ModelScope (Wang et al., 2023), MagicVideo (Zhou et al., 2022) and Stable Video Diffusion (Blattmann et al., 2023a). There are also works combining convolutional temporal layers with 3D U-Net (Blattmann et al., 2023b). Although convolutions are effective in image modeling, diverse and global relations in time dimensions are too complex limited for their local receptive field (Li et al., 2023). To additionally capture intricate global dependencies, we make attempts by proposing MT-VAE.

3 Methods

The training of PredLDM is consisted of two stages. In the first stage, PredLDM learns the distribution of signals with MT-VAE to compress spatiotemporal sequences into latent vectors to model intricate global

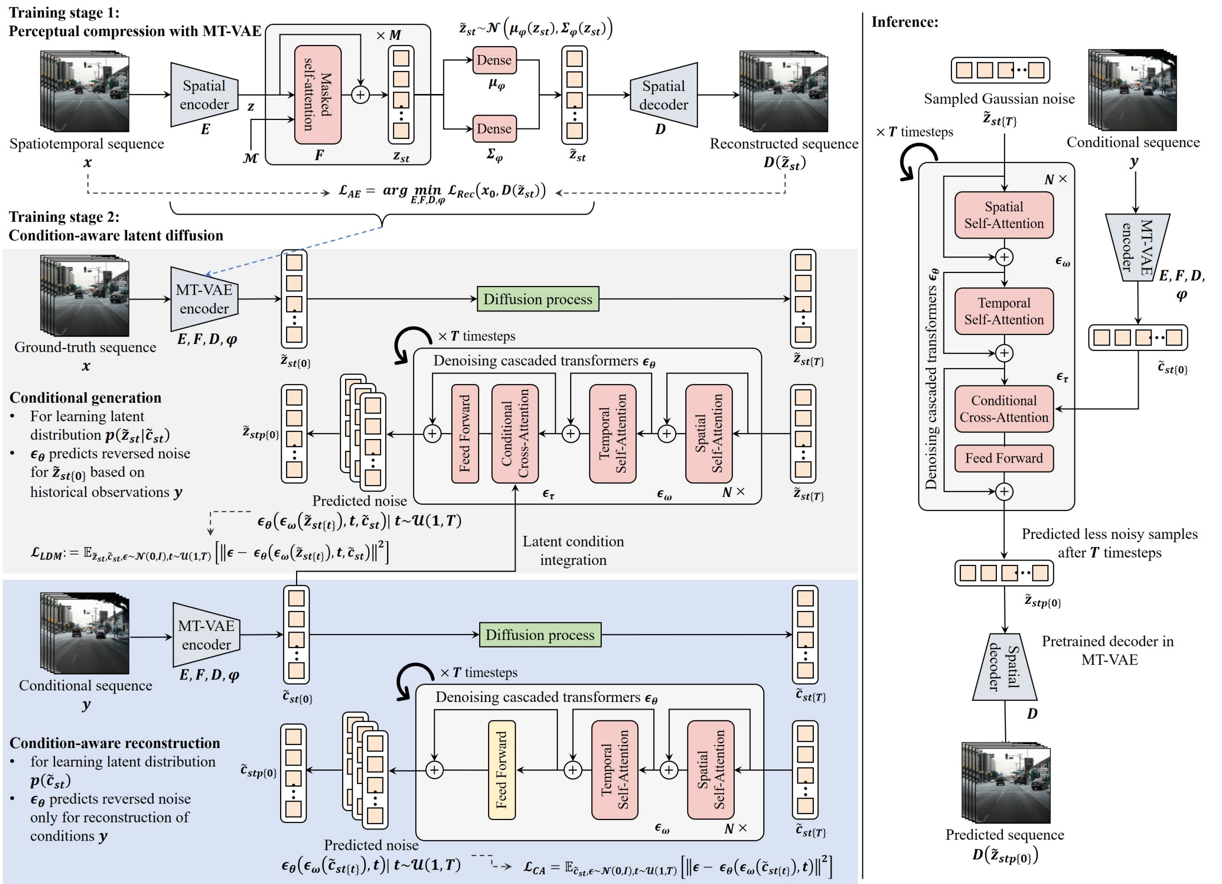

Figure 1: The pipeline of PredLDM. The training is comprised of two stages. In the first stage, MT-VAE is exploited to compress the spatiotemporal sequences into latent space. In the second stage, CA-LDM contains the learning of both conditional generation and condition-aware reconstruction. For inferencing, the Gaussian noise is sampled and less noisy latent vectors are predicted conditioned on the latent historical observations. The sampled latent vectors are then fed into the decoder for output.

coherence. In the second stage, CA-LDM is designed for comprehensive understanding of historical spatiotemporal content in conditions. To inference, a latent noise is randomly sampled and denoised with trained LDMs conditioned on historical embeddings. Predicted latent vectors are finally fed to the decoder for future content.

3.1 Background on Latent Diffusion Models

Diffusion formulation. Denoising diffusion probabilistic models (Ho et al., 2020) simulate a data distribution $x \sim p_{data}(x)$ by corrupting data with progressively added Gaussian noise and learning to reverse this process. The diffusion process leads corrupted data resembling pure noise by gradually adding Gaussian noise in a serious of timesteps, along with the sampled noisy x_t at timestep t,

$$q(x_t|x_{t-1}) = \mathcal{N}\big(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\big), \tag{1}$$

where $\{\beta_t\}_{t=1}^{T}$ are a set of linearly increasing hyperparameters with the predefined variance schedule (He et al., 2022), T denotes the number of diffusion steps and $\mathcal{N}$ refers to the normal distribution. The denoising process reverses the above diffusion process to predict less noisy x_{t-1} iteratively,

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\big), \tag{2}$$

where μ_θ and Σ_θ are accomplished by a parameterized denoising model ϵ_θ with learnable parameters θ. Specifically, $\epsilon_\theta(x_t, t)$ is trained to predict the noise at each step of the diffusion process by minimizing the difference between the actual noise and predicted ones,

$$\mathcal{L}_{DM} = \mathbb{E}_{x_0, \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}), t\sim\mathcal{U}(1,T)}[\|\epsilon - \epsilon_\theta(x_t, t)\|^2], \tag{3}$$

where $\mathcal{U}(1, T)$ refers uniformly sampling from $\{1, \cdots, T\}$.

Latent diffusion models. LDMs (Rombach et al., 2022) are efficient variants of DMs by operating in latent space. This process begins with a pretrained variational encoder $E\colon x \to z$, compressing the input image $x \sim p_{data}(x)$ into latent representations $z \sim E(x)$. Similar as Equation 1 and Equation 2 with z,

$$\mathcal{L}_{LDM} = \mathbb{E}_{E(x_0), \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}), t\sim\mathcal{U}(1,T)}[\|\epsilon - \epsilon_\theta(z_t, t)\|^2]. \tag{4}$$

3.2 Perceptual Compression with MT-VAE

In the first training stage, we compress spatiotemporal sequences with MT-VAE. The model structure of MT-VAE is provided as Figure 1. Given a sequence $x_0{\sim}p_{data}(x_0)$, $x_0 \in \mathbb{R}^{L\times H\times W\times C}$, where L, H, W and C are the temporal length, height, width and channel number respectively, the spatial convolutional encoder E encodes x_0 into latent vectors $z_0 = E(x_0)$. z_0 is taken by masked temporal self-attention modules F, extracting intricate coherent temporal reliance,

$$z_{st} = mAtt\left(Norm(z_s)\right) + z_s, z_s = (E(x_0) + U), \tag{5}$$

$$z_{st} := FeedForward\big(Norm(z_{st})\big) + z_{st}, \tag{6}$$

where U is the positional embedding obtained by convolutions of input. Insipred by scaled dot-product attention (Vaswani, 2017) and masked modeling (Xie et al., 2022; Cheng et al., 2022), $mAtt(\cdot)$ is to capture complex global reliance after the layer normalization. Multi-head mechanism (Vaswani, 2017) is used to project representations into subspaces calculated by different attention heads, the number of heads is denoted by H. The process of $mAtt(\cdot)$ is defined as below, assuming that the normalized features of z_s, $Norm(z_s)$ is z_{in}, we have

$$mAtt(z_{in}) = mAtt(Q, K, V, \mathcal{M}), \tag{7}$$

$$\left(Q^{(i)}, K^{(i)}, V^{(i)}\right) = z_{in}\left(W^{(Q,i)}, W^{(K,i)}, W^{(V,i)}\right) \tag{8}$$

$$z_{out}^{(i)} = softmax\left(\frac{Q^{(i)}K^{(i)T}}{\sqrt{d_k}} + \mathcal{M}\right)V^{(i)}, \tag{9}$$

$$where \quad \mathcal{M}_j = \begin{cases} 0, & if \quad md(j) = 1, \\ -\infty, & otherwise, \end{cases} \tag{10}$$

$$mAtt(Q, K, V) = Concat\left(z_{out}^{(1)}, \cdots, z_{out}^{(H)}\right)W^O, \tag{11}$$

where Q, K and V are queries, keys and values of vectors for dot production (Vaswani, 2017). $W^{(Q,i)}$, $W^{(K,i)}$ and $W^{(V,i)}$ are the parameters of linear operations for i-th head to control the weights of Q, K and V respectively, $i = 1, \cdots, H$. Here, $md(j) \in \{0, 1\}^L$ is the random binarized output with the masking ratio r of the same size as temporal length. $\mathcal{M}_j$ indicates all zero or all negative infinite matrices corresponding to time location j. The results from different attention heads are concatenated and projected back into representation space through the weight matrix W^O. Here the compressed feature z_{st} is accessed. The mean vectors $\mu_\varphi(z_{st})$ and variance vectors $\Sigma_\varphi(z_{st})$ are predicted with learnable parameters φ, which are implemented by two dense layers. The sampled latent features from the Gaussian distribution $\tilde{z}_{st} \sim \mathcal{N}\left(\mu_\varphi(z_{st}), \Sigma_\varphi(z_{st})\right)$ are the final compressed features used for the decoder D to reconstruct x_0. D is accomplished by the cascaded convolutional layers and $D(\tilde{z}_{st})$ is expected to minimize the difference between the predicted distributions and the real data $p_{data}(x_0)$. The training objective is the reconstruction loss with a pixel-level mean-squared error (MSE) and a perceptual loss (Johnson et al., 2016; Ni et al., 2023).

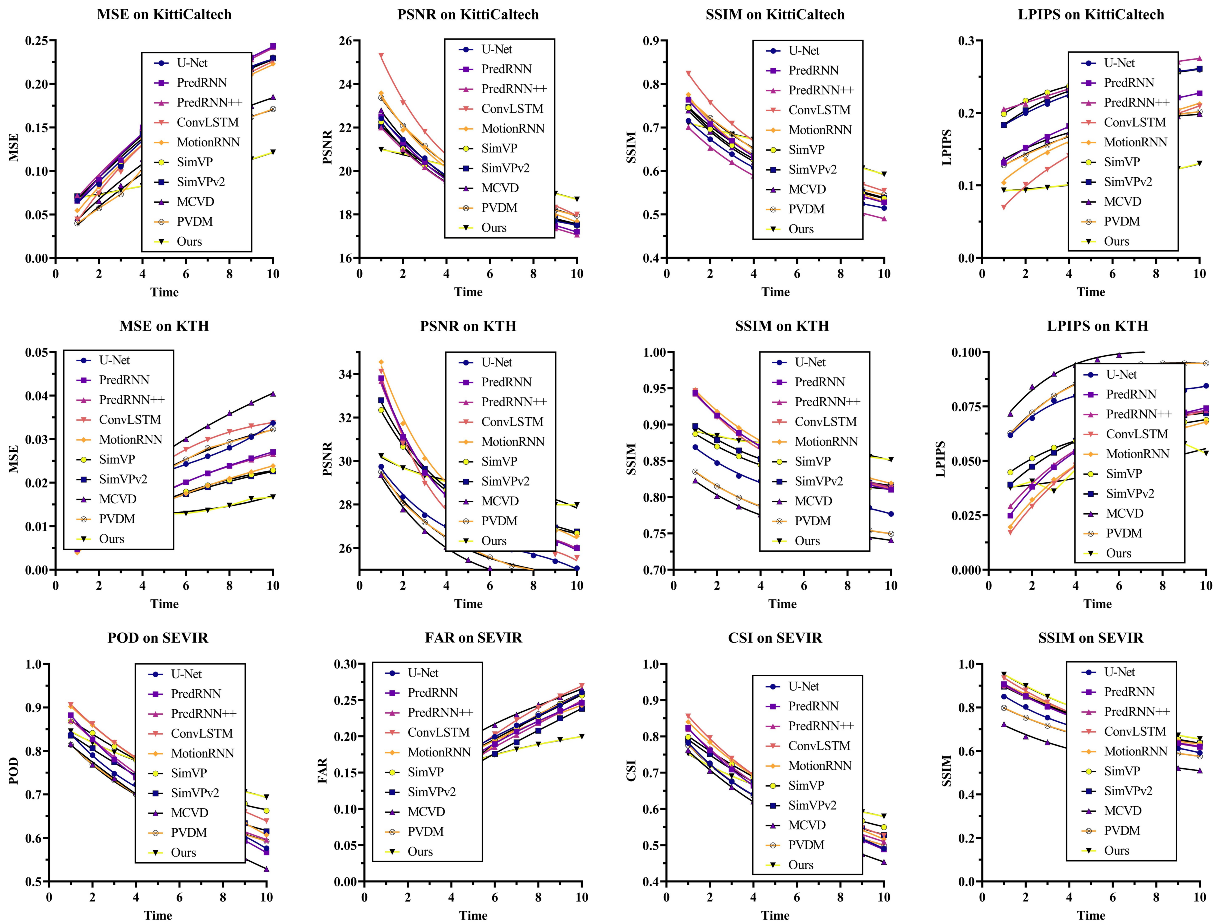

Figure 2: Temporal analysis of predictive models. Statistics at each time point is provided. Results in rows are from KittiCaltech, KTH and SEVIR respectively.

3.3 Condition-aware Latent Diffusion

In the second stage, it is expected to train a denoising network to predict less noisy samples from latent noise conditioned on historical embeddings (Sohn et al., 2015; Rombach et al., 2022). Given a sequence of future observations $x_0 \sim p_{data}(x_0)$ and its corresponding historical observations as conditions $y_0 \sim p_{data}(y_0)$, the trained MT-VAE compresses them into the latent features respectively as $\tilde{z}_{st}$ and $\tilde{c}_{st}$, $\tilde{z}_{st} \sim \mathcal{N}\left(\mu_\varphi(z_{st}), \Sigma_\varphi(z_{st})\right)$, $z_{st} = F(E(x_0))$ and $\tilde{c}_{st} \sim \mathcal{N}\left(\mu_\varphi(c_{st}), \Sigma_\varphi(c_{st})\right)$, $c_{st} = F(E(y_0))$. Our CA-LDM is trained to simultaneously learn the denoising diffusion process of (i) conditional generation on the distributions $p(\tilde{z}_{st}|\tilde{c}_{st})$ and (ii) reconstruction of the conditions on the distribution $p(\tilde{c}_{st})$, as in Figure 1.

Conditional generation. For learning $p(\tilde{z}_{st}|\tilde{c}_{st})$, the diffusion process progressively adds Gaussian noise onto $\tilde{z}_{st}$ until it resembles pure noise along with the sampled noisy $\tilde{z}_{st\{t\}}$ at timestep t. The denoising process reverses the diffusion process iteratively to approach the original latent samples $\tilde{z}_{st\{0\}}$. Instead of using time-conditional U-Net (Ronneberger et al., 2015), our denoising network ϵ_θ is inspired by DiTs (Ma et al., 2024). It is consisted of the spatiotemporal self-attentions ϵ_ω and cross-attentions ϵ_τ, where ω and τ are learnable parameters. This denoising neural network is trained by minimizing the difference between the actual noise and predicted ones,

$$\mathcal{L}_{LDM} := \mathbb{E}_{\tilde{z}_{st}, \tilde{c}_{st}, \epsilon, t}\left[\left\|\epsilon - \epsilon_\theta(\epsilon_\omega(\tilde{z}_{st\{t\}}), t, \tilde{c}_{st})\right\|^2\right], \tag{12}$$

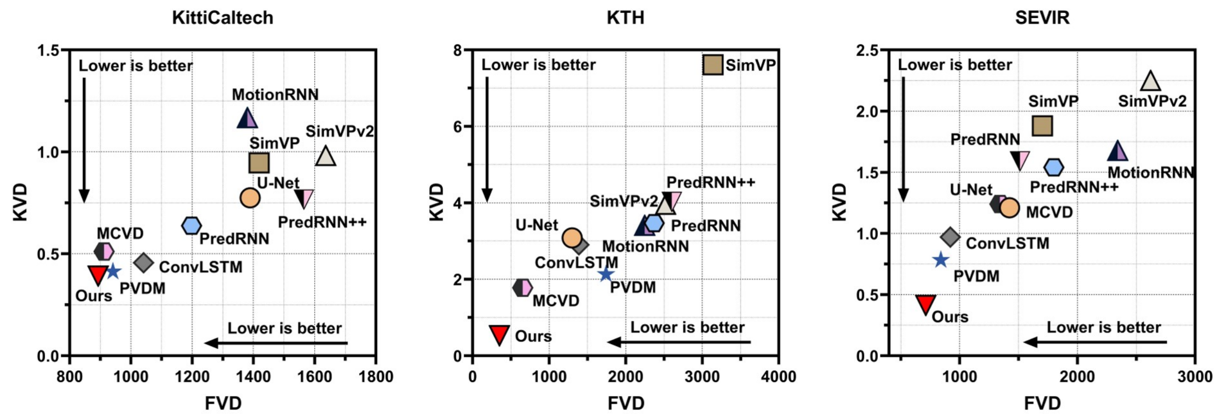

Figure 3: Plots of KVD/FVD distance scores from existing models and ours on three datasets. Distance scores between distributions of ground-truth data and predictions from predictive models are visualized. Both metrics are the lower the better.

where $t \sim \mathcal{U}(1, T)$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The condition latent features $\tilde{c}_{st}$ is integrated to the intermediate features of ϵ_θ by the cross-attention layers ϵ_τ.

Condition-aware reconstruction. For learning $p(\tilde{c}_{st})$, the diffusion process is additionally conducted onto $\tilde{c}_{st}$ with the sampled noisy $\tilde{c}_{st\{t\}}$ at timestep t. The denoising phase is operated to reverse the noisy samples to the original distribution $p(\tilde{c}_{st})$. The training objective of this branch is to progressively reconstruct the conditions by minimizing the difference between the actual noise and predicted ones from the denoising network with Siamese spatiotemporal self-attentions ϵ_ω in ϵ_θ yet without cross-attentions,

$$\mathcal{L}_{CA} = \mathbb{E}_{\tilde{c}_{st}, \epsilon, t}\left[\left\|\epsilon - \epsilon_\theta(\epsilon_\omega(\tilde{c}_{st\{t\}}), t)\right\|^2\right]. \tag{13}$$

The overall learning objective of CA-LDM is the combination of $\mathcal{L}_{LDM}$ and $\mathcal{L}_{CA}$,

$$\mathcal{L}_{CA-LDM} = \mathcal{L}_{LDM} + \mathcal{L}_{CA}. \tag{14}$$

3.4 Inference

As in Figure 1, to inference the expected spatiotemporal sequence x_0 with the condition y_0, a Gaussian noise z_T is sampled in the latent space and the condition is compressed by MT-VAE as $\tilde{c}_{st}$. The denoising network ϵ_θ in CA-LDM is used to predict the less noisy samples $z_{0:T-1}$ within the predefined T timesteps, while the latent condition vectors $\tilde{c}_{st}$ are fused across cross-attention layers into the intermediate features of ϵ_θ to accomplish conditional controlling. The less noisy sample can be predicted by $z_{t-1} \sim \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t))$. The latent vectors z_0 can be approached after T timesteps denoising. Finally, the decoder in MT-VAE decodes z_0 back to the pixel space, resulting predictions as close as possible to ground-truth x_0.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. Datasets in this study include KittiCaltech (Geiger et al., 2013), KTH (Schuldt et al., 2004) and SEVIR (Veillette et al., 2020). (i) KittiCaltech is a driving-scene dataset, comprising a curated collection of high-quality images. The ability to predict the future dynamics of this scenario is paramount for the advancement of autonomous driving technology and dynamic comprehension, containing $127,271$ frames in total, with $74,833$ frames for training and $52,438$ frames for testing. (ii) KTH stands as a benchmark in the field of human action recognition and prediction. It encompasses diverse image sequences depicting a variety of human activities. This dataset is comprised of $51,360$ frames with $20,420$ frames

Table 1: Comparison between existing models and PredLDM on KittiCaltech, KTH as well as SEVIR datasets. ↑ indicates the higher the better, whereas ↓ is the opposite. The best results are marked as bold and the second best ones are marked as underline.

Models	KittiCaltech			KTH			SEVIR		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	POD ↑	FAR ↓	CSI ↑
U-Net	19.34	0.591	0.232	26.75	0.813	0.079	0.691	0.185	0.612
PredRNN	19.19	0.616	0.190	28.33	0.860	0.057	0.704	0.174	0.629
PredRNN++	18.99	0.571	0.244	28.33	0.861	0.058	0.714	0.171	0.638
ConvLSTM	**20.46**	<u>0.652</u>	<u>0.154</u>	27.87	0.859	0.052	<u>0.754</u>	0.182	<u>0.660</u>
MotionRNN	19.79	0.621	0.170	<u>28.94</u>	<u>0.868</u>	<u>0.051</u>	0.742	0.177	0.654
SimVP	19.19	0.614	0.239	28.47	0.838	0.060	0.752	0.184	0.657
SimVPv2	19.29	0.620	0.234	28.64	0.847	0.061	0.713	<u>0.163</u>	0.641
MCVD	19.41	0.607	0.177	25.78	0.770	0.094	0.659	0.197	0.583
PVDM	19.94	0.631	0.174	26.25	0.781	0.086	0.681	0.173	0.610
PredLDM	<u>19.86</u>	**0.653**	**0.107**	**28.94**	**0.871**	**0.045**	**0.760**	**0.148**	**0.672**

Table 2: Influence of different settings of autoencoders. ↑ indicates the higher the better, whereas ↓ is the opposite. The best results are marked as bold.

Autoencoders	KittiCaltech		KTH	
	SSIM ↑	LPIPS ↓	SSIM ↑	LPIPS ↓
3D VAEs	0.630	0.157	0.861	0.060
2D VAEs + 1D Convs	0.647	0.121	0.869	0.045
MT-VAE (Ours):	**0.653**	**0.107**	**0.871**	**0.045**

Table 3: Influence of condition-aware latent diffusion. ↑ indicates the higher the better, whereas ↓ is the opposite. The best results are marked as bold.

LDMs	KittiCaltech		KTH	
	SSIM ↑	LPIPS ↓	SSIM ↑	LPIPS ↓
LDM $\mathcal{L}_{LDM}$	0.622	0.184	0.858	0.067
CA-LDM (Ours): LDM $\mathcal{L}_{LDM} + \mathcal{L}_{CA}$	**0.653**	**0.107**	**0.871**	**0.045**

for training and $30,940$ frames for testing. (iii) SEVIR has been curated in the realms of weather sensing and short-term forecasting, comprising thousands of weather events in multipe sensor modalities. We use vertically integrated liquid (VIL) data with a 5-minute interval, and 1 km spatial resolutions. They are stored as integers ranging from 0 to 254, with a value of 255 indicating missing data. In our experiments, all frames are processed as 128×128 resolutions. The temporal length of input is uniformly 10 frames and output length is also 10 frames. For evaluating on KittiCaltech and KTH, the metrics contain MSE, PSNR, SSIM (Jin et al., 2020; Wang et al., 2004) and LPIPS (Zhang et al., 2018). For evaluating weather patterns on SEVIR, we use event-level short-term prediction metrics and the image quality assessment metric, including POD (Veillette et al., 2020), FAR (Veillette et al., 2020), CSI (Schaefer, 1990) and SSIM. More details can be accessed as in Supplementary Section A.1.

Implementation details. PredLDM is trained in two stages. In the first stage, spatiotemporal sequences are autoencoded by MT-VAE. The masking ratio r is set as 0.6. In the second stage, trained parameters in

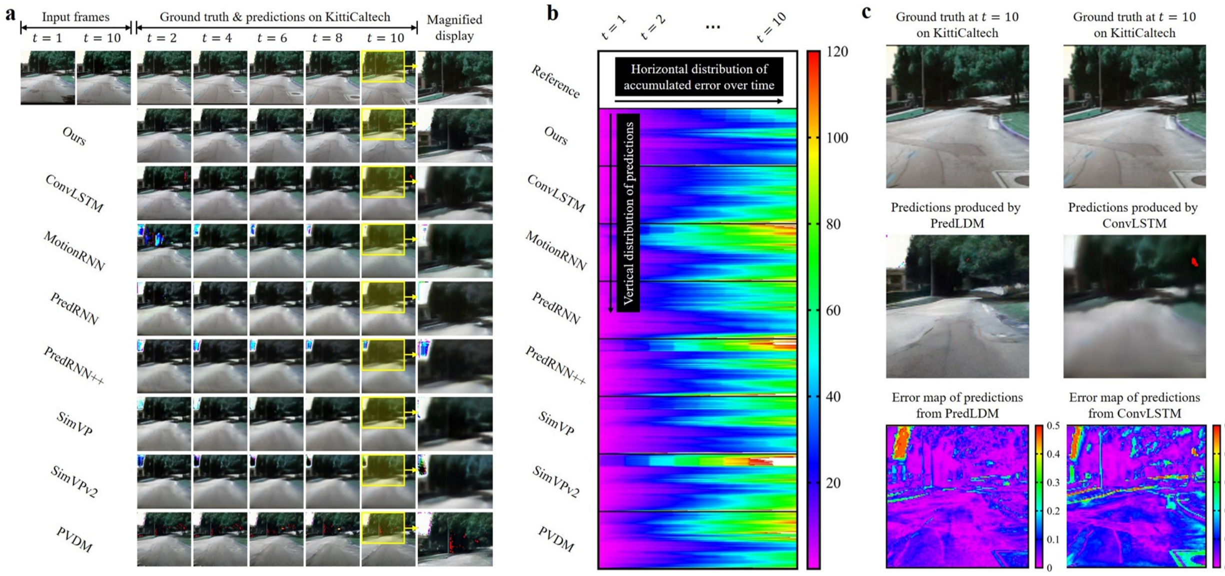

Figure 4: Predictive error analysis. **a,** A challenging case on KittiClatech is visualized. **b,** Analysis on error of predictions accumulated with time. We calculate the accumulated error along with x-axis, the distribution should be all blue if there is no error accumulated. **c,** Error maps of the last frame. The highlighted value indicates the largest error.

MT-VAE including E, F and φ are used to project data into latent space. The denoising model is trained to predict less noisy samples by the linear combination of $\mathcal{L}_{LDM}$ and $\mathcal{L}_{CA}$. More details and hyper parameters implementing our PredLDM are available as in Supplementary Section A.1.

4.2 Results with Comparison to Existing Models

Baselines in our experiments include classical encoder-forecaster architectures U-Net (Ronneberger et al., 2015), PredRNN (Wang et al., 2017), PredRNN++ (Wang et al., 2018), ConvLSTM (Shi et al., 2015), MotionRNN (Wu et al., 2021b), SimVP (Gao et al., 2022), SimVPv2 (Tan et al., 2023) as well as diffusion-based probabilistic generation architectures MCVD (Voleti et al., 2022) and PVDM (Yu et al., 2023). As in Table 1, PredLDM almost achieves best scores in all metrics on KittiCaltech and KTH, except in PSNR. The most shining one is LPIPS which resembles the perceptual evaluation alike to human, where 30.5% improvement is achieved by PredLDM on KittiCaltech and 11.8% improvement is made on KTH. Results on weather nowcasting show large margin in scores of PredLDM over other models.

4.3 Temporal Analysis

For observing temporal detailed performance, results are compiled by performance at each time point as in Figure 2. It is evident that the descending curve decreases slowly for PredLDM, while the degradation of other models is quite fast. This implies that PredLDM is capable of dealing with complex global temporal variations. This phenomenon is more prominent in LPIPS on KittiCaltech and KTH.

4.4 Experiments on KVD/FVD Comparison

For further analyzing the temporal retention ability of models, we calculate Fréchet Video Distance (FVD) and Kernel Video Distance (KVD) (Unterthiner et al., 2018) between predictions and groundtruth on KittiCaltech, KTH and SEVIR datasets. Different from frame-level metrics, these two metrics measure the similarity of sequential distributions. As in Figure 3, results indicate better retention abilities of visual quality in temporal dimensions.

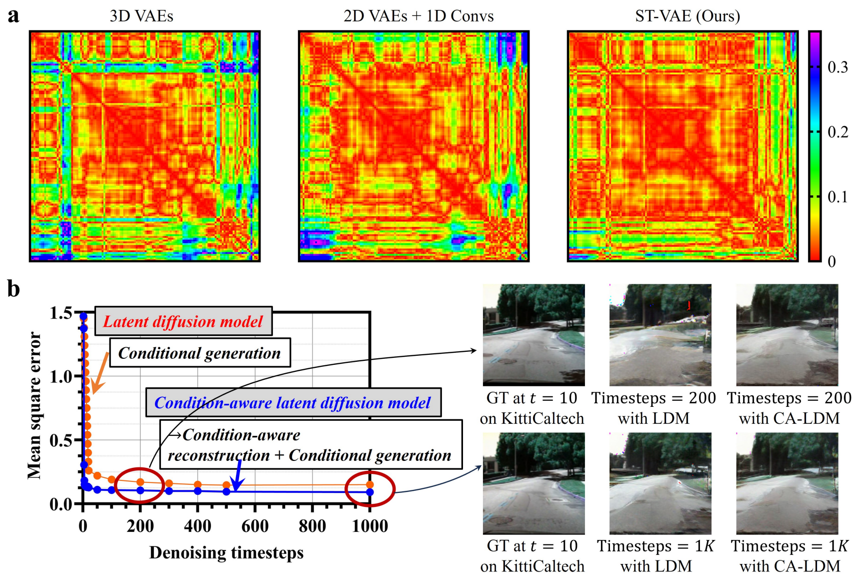

Figure 5: Ablation study. **a,** Heatmaps of correlation between synthetic data v.s. real data for MT-VAE. **b,** Influence of condition-aware latent diffusion. Performance in denoising processes is plotted on the left and the decoded predictions are on the right.

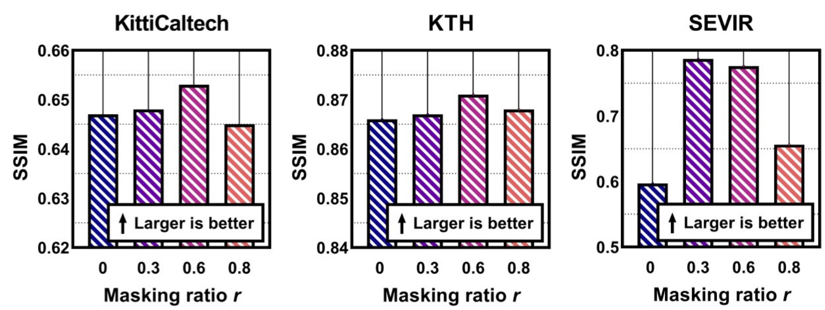

Figure 6: Influence of masking ratios on forecasting performance. SSIM scores with different masking ratios of PredLDM on KittiCaltech, KTH and SEVIR are reported.

4.5 Predictive Error Analysis

For analyzing the predictive error between existing competitive implementations and ours, we select a challenging sample when a car is driving through the corner. As in Figure 4, the visual appearance from PredLDM shares quite similar details as the ground truth while others show evident visual difference. For visualizing the accumulated error over time on these models, we count the cumulative values of all absolute errors along the x-axis direction through time. The plot shows that the accumulated error over time is smaller for predictions from PredLDM than others. For the error map calculated from the last predicted frame, the error of the results is very limited from ours. It can be observed that the visual quality and accumulative error are evidently improved by our model.

4.6 Ablation Study

For organizing the ablation study, we firstly compile the performance on different settings of autoencoders including 3D VAEs, 2D VAEs + 1D convolutions and ours MT-VAE, as in Table 2. It can be seen that the setting of 2D VAEs with 1D convolutional attention is more competitive than 3D VAEs, consistent with

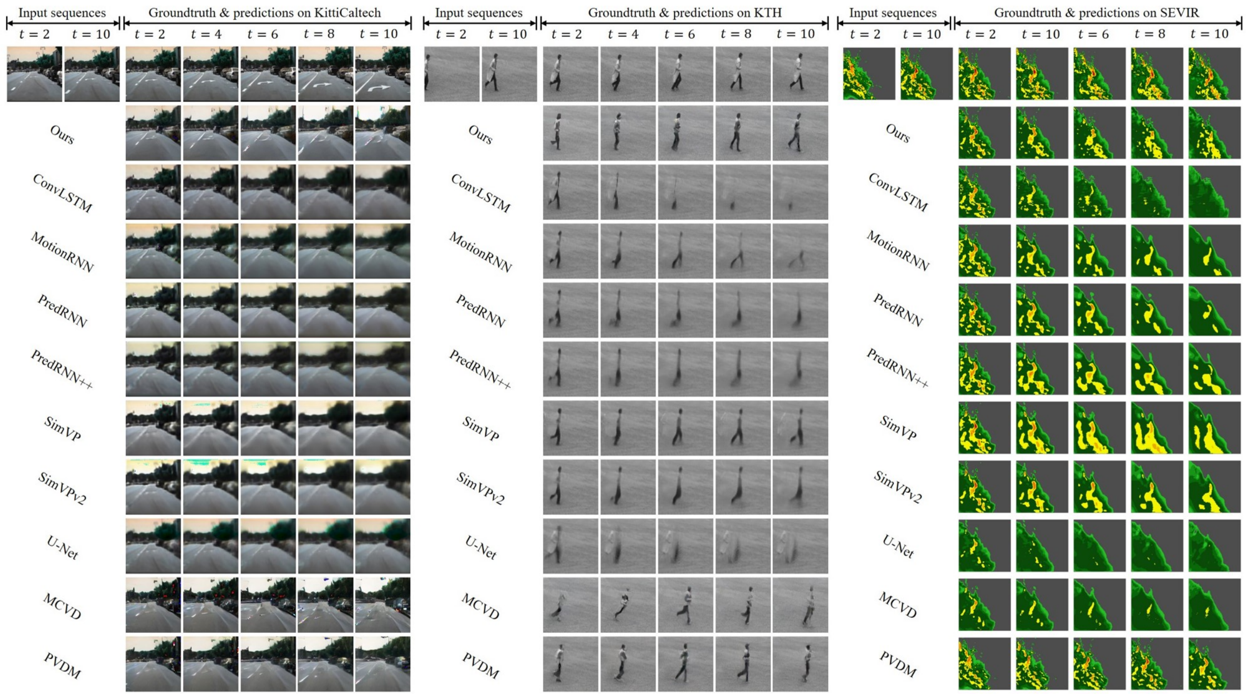

Figure 7: Challenging cases on three datasets. Visualized challenging cases on KittiCaltech, KTH and SEVIR are presented. The first row refers to input and ground-truth spatiotemporal sequences, following rows indicate predictions.

existing research (Tran et al., 2018). Our setting of MT-VAE is better than this competitive one. Then we report the influence of condition-aware latent diffusion as in Table 3, showing that the condition-aware reconstruction-based loss is beneficial.

Further study is conducted as Figure 5. Correlations are analyzed from settings of autoencoders in Table 2. It shows that our MT-VAE is better for generating realistic data, where the correlation-based distributions of ours indicate the most similar behavior as realistic structures. The mean square error of sampled results during the denoising process shows that the addition of condition-aware constraint brings lower error. From the decoded visual appearance, CA-LDM can produce more realistic visual quality. More analysis is available in Supplementary Section A.2.

4.7 Influence on Masking Ratios

For investigating the influence of the masking ratio r on forecasting performance on three datasets, different settings of masking ratios are experimented as in Figure 6. More results are provided as in Supplementary Section A.3. Results show effectiveness of masked modeling and the setting of masking ratio is preferred as 0.6 in our experiments.

4.8 Case Study

Predictions of existing models and ours PredLDM on challenging cases of three datasets can be accessed in Figure 7. It can be seen that the predictions form PredLDM not only show the realistic visual appearance, but also share the most similar movement as the ground-truth sequences. More examples are available in Supplementary Section A.4.

5 Conclusion

In this study, we propose a spatiotemporal predictive model with LDMs called PredLDM, towards predicting the accurate and realistic future. Under consideraton of intricate global coherence and comprehensive history understanding, corresponding designs are made. (i) MT-VAE is proposed to compress intricate global coherent spatiotemporal latent representations with the combination of transformers with masked attention and convolutional VAEs. (ii) CA-LDM is proposed by learning distributions of both conditional generation and condition-aware reconstruction, to comprehensively understand conditions which are spatiotemporal sequences with diverse and complex context. Through extensive experiments on multiple scenarios, PredLDM shows accurate performance and realistic appearance in predictions, revealing promising potential in future research and applications.

References

Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*, 2021. 1, 3

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 3

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a. 3

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023b. 3

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022. 2, 5

Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019. 1, 3

Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1472–1480, 2024. 3

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G Derpanis, and Bjorn Ommer. Stochastic image-to-video synthesis using cinns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3742–3753, 2021. 1, 3

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021. 3

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016. 1

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3170–3180, 2022. 1, 2, 9

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2, 7, 17

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 1, 2, 3, 4

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a. 3

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022b. 1, 2, 3

Yanping Huang and Rajesh PN Rao. Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5):580–593, 2011. 2

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*, pp. 4554–4563, 2020. 8

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016. 5

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 3

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 17

Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1811–1820, 2019. 1

Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018. 1

Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12581–12600, 2023. 2, 3

Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 3

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023. 3

Yutian Liu, Soora Rasouli, Melvin Wong, Tao Feng, and Tianjin Huang. Rt-gcn: Gaussian-based spatiotemporal graph convolutional network for robust traffic prediction. *Information Fusion*, 102:102078, 2024. 1

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 3, 6

Midjourney. Midjourney. https://www.midjourney.co, 2023. 3

Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18444–18455, 2023. 5

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 3

Marc Oliu, Javier Selva, and Sergio Escalera. Folded recurrent neural networks for future video prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 716–731, 2018. 1, 2

OpenAI. Dalle-2. https://openai.com/dall-e-2, 2023. 3

OpenAI. Sora. https://openai.com/index/sora/, 2024. 3

Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2806–2826, 2020. 1, 2, 3

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. 3, 18

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014. 2

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999. 2

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 3

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 1, 3, 5, 6, 17

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015. 3, 6, 9

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

Joseph T Schaefer. The critical success index as an indicator of warning skill. *Weather and forecasting*, 5 (4):570–575, 1990. 8

Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pp. 32–36. IEEE, 2004. 2, 7, 17

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015. 1, 2, 3, 9

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 3

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 6

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pp. 843–852. PMLR, 2015. 2

Mingzhen Sun, Weining Wang, Xinxin Zhu, and Jing Liu. Moso: Decomposing motion, scene and object for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18727–18737, 2023. 1, 2

Cheng Tan, Siyuan Li, Zhangyang Gao, Wenfei Guan, Zedong Wang, Zicheng Liu, Lirong Wu, and Stan Z Li. Openstl: A comprehensive benchmark of spatio-temporal predictive learning. *Advances in Neural Information Processing Systems*, 36:69819–69831, 2023. 1, 2, 9

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015. 3

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018. 3, 11

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1526–1535, 2018. 1, 3

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 9

Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, Karsten Kreis, et al. Lion: Latent point diffusion models for 3d shape generation. *Advances in Neural Information Processing Systems*, 35: 10021–10039, 2022. 3

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 5

Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in Neural Information Processing Systems*, 33: 22009–22019, 2020. 2, 7, 8, 17

Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *arXiv preprint arXiv:1706.08033*, 2017. 1, 2

Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022. 3, 9

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 3

Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30, 2017. 1, 2, 9

Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and S Yu Philip. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *International conference on machine learning*, pp. 5123–5132. PMLR, 2018. 1, 2, 9

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 8

Bohan Wu, Suraj Nair, Roberto Martin-Martin, Li Fei-Fei, and Chelsea Finn. Greedy hierarchical variational autoencoders for large-scale video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2318–2328, 2021a. 1, 3

Haixu Wu, Zhiyu Yao, Jianmin Wang, and Mingsheng Long. Motionrnn: A flexible model for video prediction with spacetime-varying motions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15435–15444, 2021b. 1, 2, 9

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9653–9663, 2022. 2, 5

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18456–18466, 2023. 9

Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. *arXiv preprint arXiv:2310.09656*, 2023a. 3

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*, pp. 586–595, 2018. 8

Yuchen Zhang, Mingsheng Long, Kaiyuan Chen, Lanxiang Xing, Ronghua Jin, Michael I Jordan, and Jianmin Wang. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619(7970):526–532, 2023b. 1

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 3

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024. 3

A Appendix

Organization of this supplementary file. In this supplementary file, we provide (i) detailed descriptions of data, hyper parameters and implementation in our experiments, as in Section A.1. (ii) Additional analysis of PredLDM especially with respect to critical designs including MT-VAE and CA-LDM as in Section A.2. (iv) More results on investigating influence of masking ratios are given in Section A.3. (v) For additional support of case study, more challenging cases with different predictive models are presented as in Section A.4.

A.1 More Details on Data and Implementation

Data. Datasets in this study include driving scene-related dataset KittiCaltech (Geiger et al., 2013), human action-related dataset KTH (Schuldt et al., 2004) and weather pattern-related dataset SEVIR (Veillette et al., 2020). (i) KittiCaltech dataset is a cornerstone in the domain of computer vision, serving as an essential resource for autonomous driving research. It comprises a curated collection of high-quality images that are vital for the understanding of driving scenarios. The ability to predict the future dynamics of these scenarios is paramount for the advancement of autonomous driving technology, rendering this dataset exceptionally valuable for research in vision field, particularly in the areas of future scenario prediction and dynamic comprehension. This dataset is meticulously organized, consisting of a total of $127,271$ frames. Within this collection, $74,833$ frames are allocated for training purposes, while $52,438$ frames are reserved for testing. This structured distribution ensures a comprehensive framework for both the development and validation of autonomous driving algorithms. (ii) KTH dataset stands as a benchmark in the field of human action recognition and prediction. It encompasses a diverse array of image sequences depicting a variety of human activities, such as walking, jogging, running, boxing, waving, and clapping, totaling six distinct categories, capturing the intricacies of different individuals performing various actions. Comprising a total of $51,360$ frames, the KTH dataset is segmented into $20,420$ frames for training and $30,940$ frames for testing. To maintain uniformity in evaluation, all frames are centrally cropped and resized to a consistent dimension of 128×128 pixels. The dataset's processing protocol specifies that the input consists of 10 frames, with the output also comprising 10 frames, ensuring a standardized framework for analysis and comparison. (iii) SEVIR dataset has been curated to accelerate research in the realms of weather sensing, avoidance and short-term forecasting. This comprehensive collection comprises thousands of weather events with each represented as a 4-hour sequence. Researchers are empowered to synthesize and harmonize diverse weather sensor data into a unified dataset through SEVIR. The dataset encompasses a variety of sensor modalities, including IR069 (infrared satellite imagery at 6.9 m), IR107 (infrared satellite imagery at 10.7 m), VIL (vertically integrated liquid), and LGHT (Lightning). In this study, we use VIL modality. The VIL data is derived from NEXRAD radar mosaics, featuring a 384×384 pixel resolution, a 5-minute interval, and a 1 km spatial resolution. The geographically and chronologically aligned imagery, depicting a spectrum of weather events including high winds, tornadoes, and hail, is captured by GEOS-16 satellites and NEXRAD weather radars. This data is publicly available in HDF files, we convert them into recordings of images. The pixel values within these spatial grids correspond to processed statistics derived from the actual sensor readings. VIL images are stored as integers ranging from 0 to 254, with a value of 255 indicating missing data. All frames have been reprocessed to a 128×128 resolution. The temporal length of input is uniformly 10 frames and output length is also 10 frames.

Implementation Details. The training of PredLDM is comprised of two stages. In the first stage, spatiotemporal sequences are autoencoded by our MT-VAE with the loss $\mathcal{L}_{AE}$. The ADAM optimizer (Kingma, 2014) with a constant learning rate of $1e-4$ is used. The batch size for training is set to be 4 and the number of total epochs is 100. The pretrained weights from image-based 2D VAEs (Rombach et al., 2022)

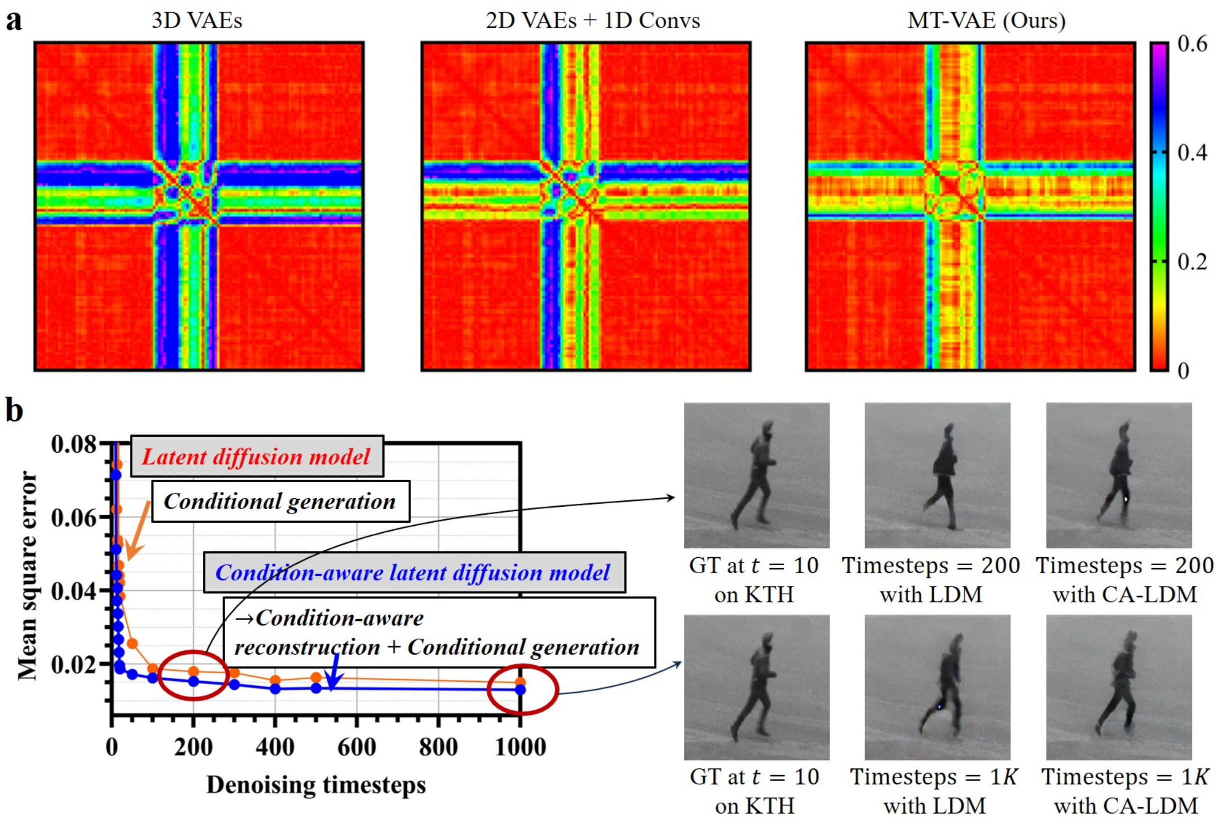

Figure 8: Additional ablation study on KTH. **a,** Heatmaps of correlation between synthetic data v.s. real data for MT-VAE. **b,** Influence of condition-aware latent diffusion. Performance in denoising processes is plotted on the left and the decoded predictions are on the right.

used for image synthesis are employed for the initialization of the convolutional encoder and decoder, where the loaded weights of E, D stay fixed during training. The parameters of the transformer F and dense layers φ are updated in this stage. No weight decaying schedule and data augmentation is used. For the designed temporal self-attention architecture, the number M of stacked attention layers is set to be 8 and the hidden dimension is 128 in this study. The masking ratio r is set as 0.6. In the second stage, the trained parameters in MT-VAE including E, F and φ are used to project data into latent space. In latent space, the denoising model is trained to predict less noisy samples by the linear combination of $\mathcal{L}_{LDM}$ and $\mathcal{L}_{CA}$. The diffusion transformer (DiT) structure (Peebles & Xie, 2023) is used for constructing the denoising model with cascaded transformers. The ADAM optimizer with the same learning rate of $1e-4$ is used. The batch size is set to be 4 and the total epochs are 100. The parameters of the denoising model are the only learnable parameters in this stage. There are also no special weight decaying and augmentation schedules used. The number N of cascaded DiTs is set to be 32 and the hidden size is 1152. These two stages are both performed on NVIDIA GeForce RTX 4090 with 24 GB $\times 4$. To inference with our PredLDM, the process is conducted and the timesteps T for denoising is 1000.

A.2 Additional Analysis on Critical Designs

For additional study of critical designs on KTH and SEVIR datasets, different settings of autoencoders including 3D VAEs, 2D VAEs + 1D convolutions and our MT-VAE are evaluated by pair-wise column correlation between synthetic data and real data. The value in heatmaps of correlation indicates the absolute divergence, where the more red area means the better correlation to real distributions. Meanwhile the performance calculated from predicted less noisy samples and ground truth during denoising processes with

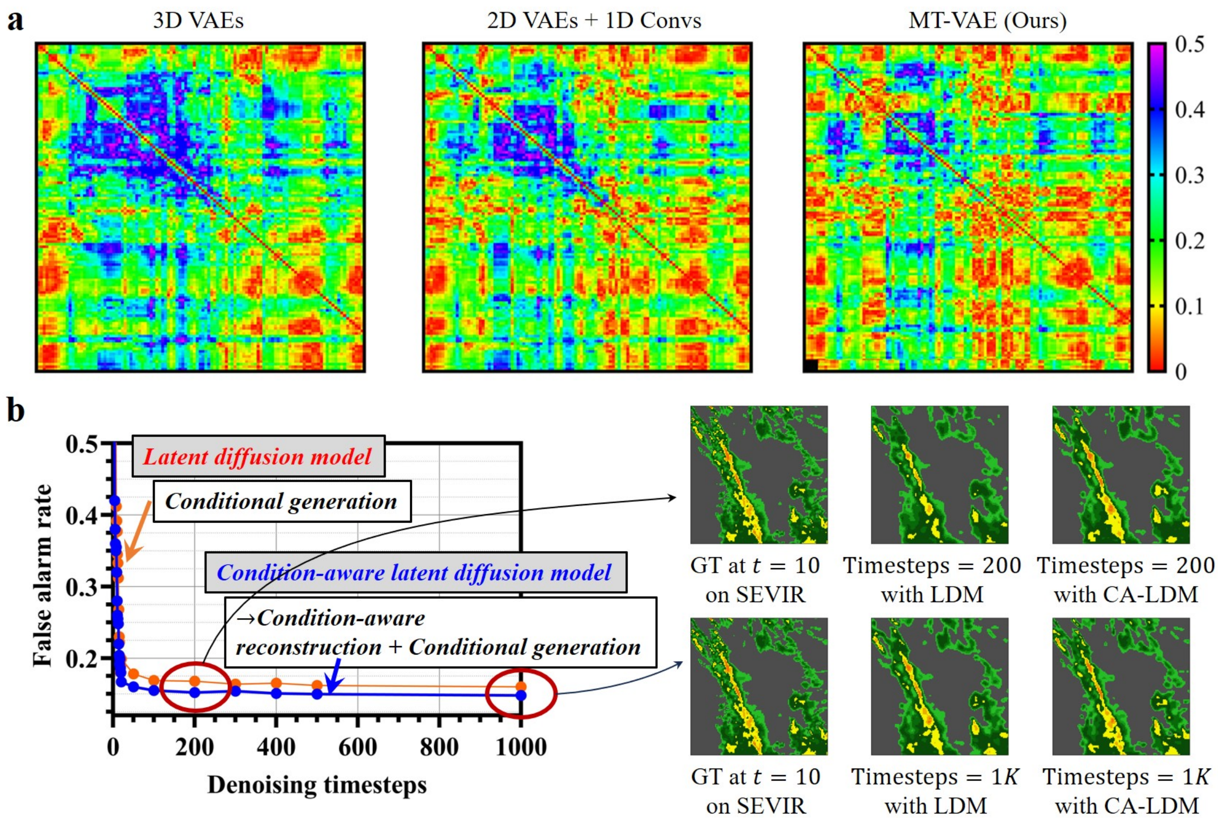

Figure 9: Additional ablation study on SEVIR. **a,** Heatmaps of correlation between synthetic data v.s. real data for MT-VAE. **b,** Influence of condition-aware latent diffusion. Performance in denoising processes is plotted on the left and the decoded predictions are on the right.

the trained denoising network with or without condition-aware reconstruction based constraints is plotted. The decoded predictions at different timesteps on these two datasets are also provided in this section.

For analysis on KTH dataset, the results can be seen as in Figure 8. As the directions of human movement in this dataset are always vertical or horizontal and close to the center of images, the heatmaps of correlation reflect this characteristic. Results show that the synthetic distributions of MT-VAE are the closest to the real distribution. The setting of 2D VAEs + 1D convolutional temporal layers follows our setting and the setting of 3D VAEs behaves not competitive. This phenomenon shows that our MT-VAE is better at compressing spatiotemporal content into latent space with realistic distributions and our setting is effective. From the mean square error of sampled results during the denoising process, it is evident that the condition-aware reconstruction of CA-LDM is beneficial as the error of predictions from CA-LDM is much lower than the trained denoising model with only the original conditional generation related constraint. For the decoded visual results from timesteps at 200 and $1,000$, the decoded visual quality from CA-LDM is better than the LDM setting and more timesteps are more conducive to producing detailed realistic visual appearance.

For analysis on SEVIR dataset, results are available as in Figure 9. The evolution of observations related to vertical liquid precipitation captured by the weather radar is distributed diversely in terms of physical geographic space, resulting the predicted dynamics difficult to be similar as ground truth, so heatmaps of correlation here are distributed in a disorderly manner. Results show that our setting of MT-VAE is still the best choice compared to other two settings, with highest relation to the distribution of real data. This again reveals that our MT-VAE is better at handling perceptual autoencoding. From the mean square error of sampled results during denoising processes, it can be seen that the false alarm rate of our predictions is much lower than predictions from the denoising model without condition-aware constraints. For the decoded

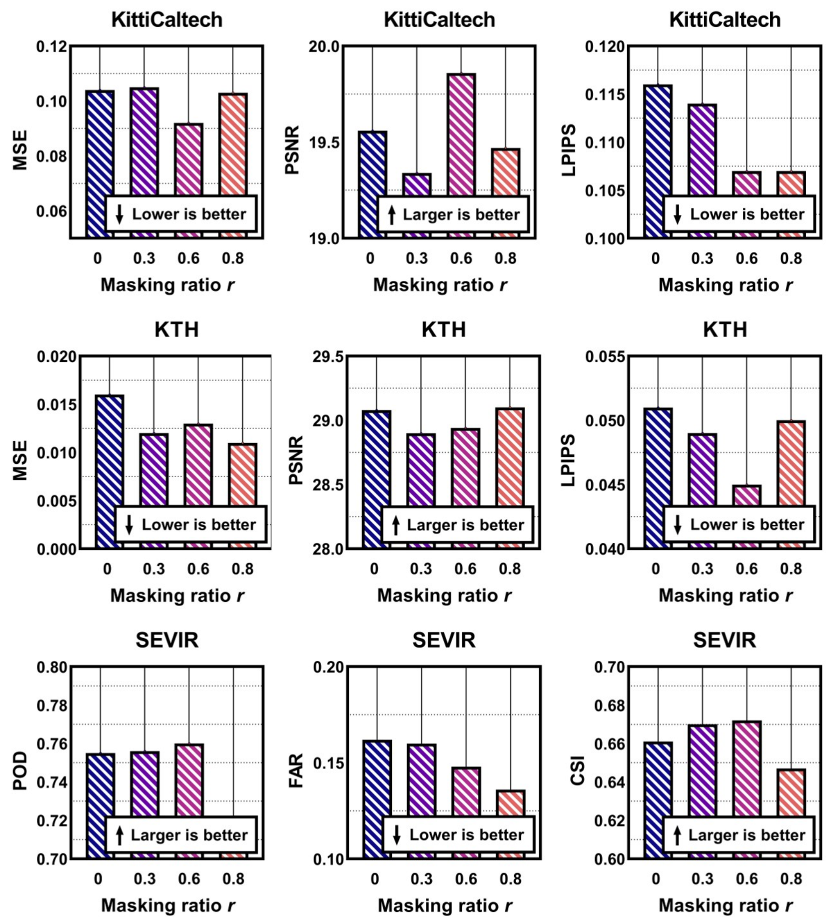

Figure 10: More results on influence of masking ratios. Remaining metrics on three datasets are presented.

visual results from different timesteps, the decoded predictions in this dataset reveal more accurate dynamics of heavy precipitation and realistic distributions of CA-LDM. Besides, realistic visual appearance with more details can be accessed by more denoising timesteps.

A.3 More Results on Influence of Masking Ratios

To supplement the results in checking influence of masked modeling, we provide more results as in Figure 10. These results show the similar phenomenon as in experiments of the manuscript that the masked attention is necessary in our pipeline and the preferred setting is 0.6.

A.4 More Challenging Cases

Additional challenging cases are visualized with predictions by predictive models on KittiCaltech, KTH and SEVIR datasets. As in Figure 11 and Figure 12, the challenging cases on KittiCaltech are presented. The predictions show the accurate modeling capability in spatiotemporal variations on car driving scenes, with realistic visual appearance compared to other models where our predictions are of high fidelity and results of other models are in more blur over time. As in Figure 13 and Figure 14, the challenging cases on KTH dataset are available. In these challenging cases, we can witness that many models are in trouble of handling these cases, where small parts of image sequences appear significant variations of motion. Predictions from existing models are highly likely to generate blur, while the predictions of PredLDM are quite attractive that they still appear realistic even to the last predicted frame. Besides, the better accuracy of predicted

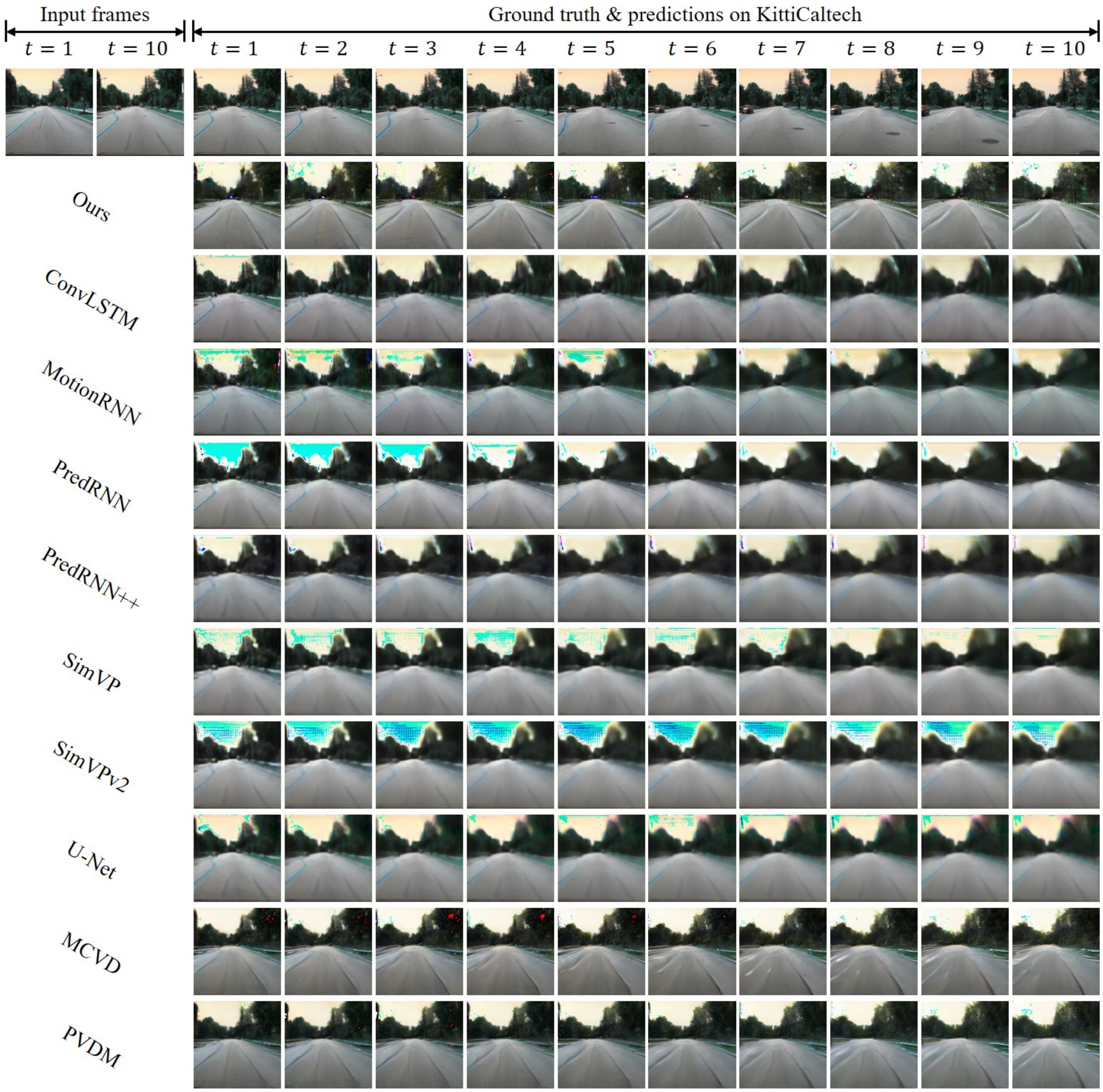

Figure 11: Challenging cases on KittiCaltech dataset. The first row refers to input spatiotemporal sequences and ground truth. The second row indicates sequences predicted by our PredLDM. Following rows are predictions from other models.

movement in these cases is evident for our model. As in Figure 15 and Figure 16, the challenging cases on SEVIR dataset are given. In weather pattern-related examples, existing predictive models are still likely to produce precipitation values with rough geophysical details, where the locations with high probabilities of rainfall are easy to omit, leading high risk for social failure preventing natural disasters. However, to results from PredLDM, predicted locations of heavy precipitation are more accurate and the overall distribution is more alike to the ground truth.

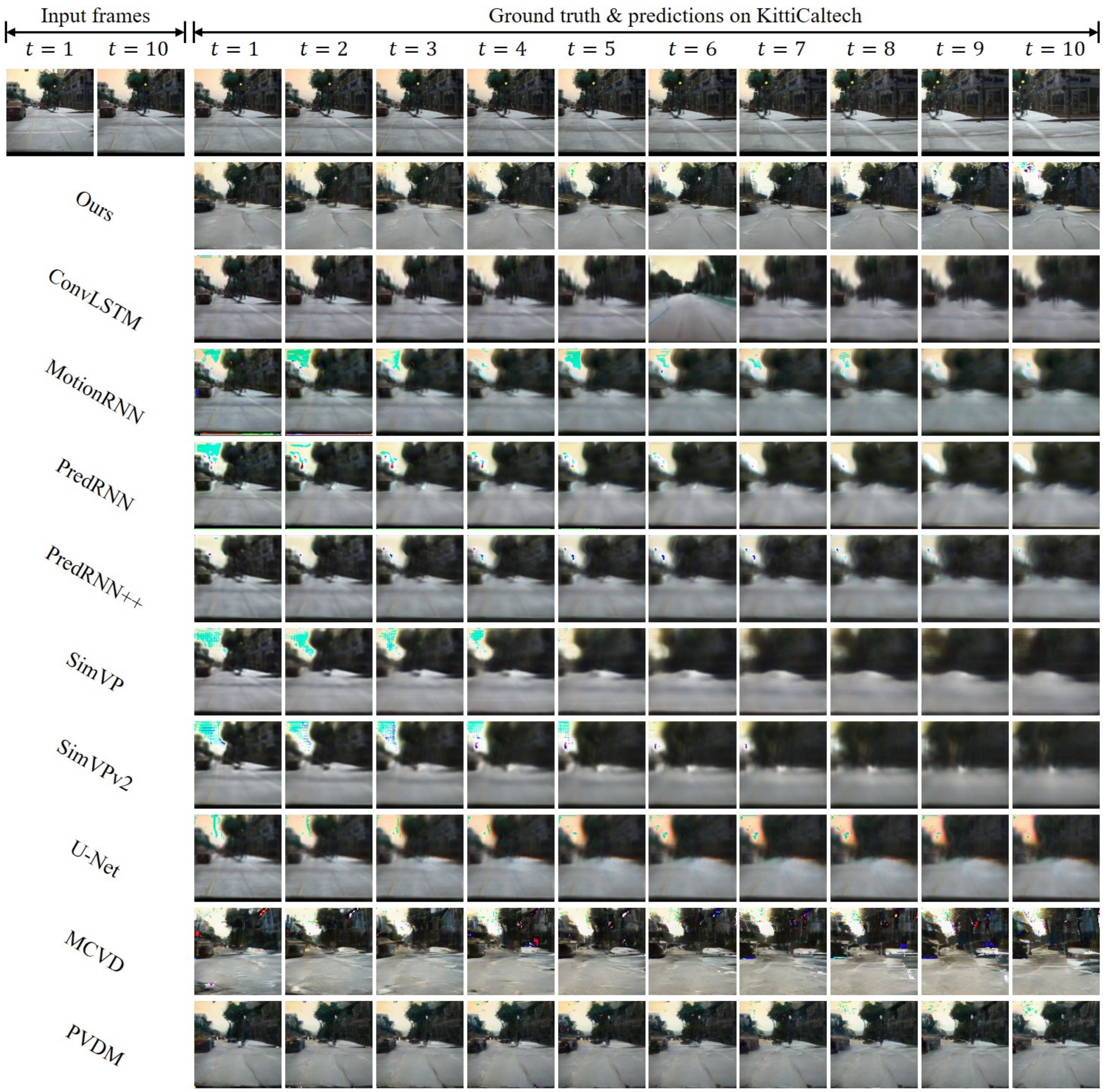

Figure 12: Challenging cases on KittiCaltech dataset. The first row refers to input spatiotemporal sequences and ground truth. The second row indicates sequences predicted by our PredLDM. Following rows are predictions from other models.

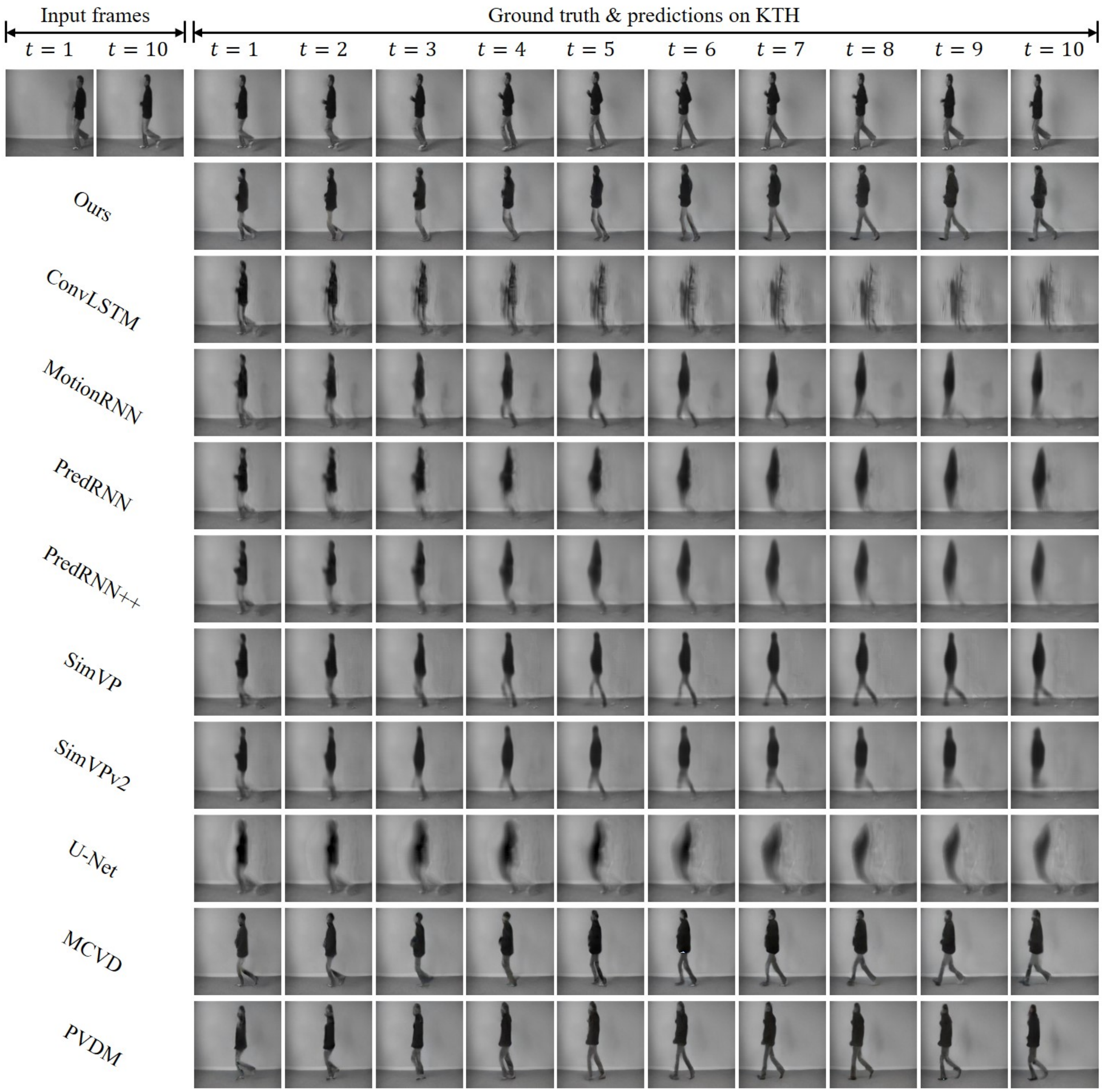

Figure 13: Challenging cases on KTH dataset. The first row refers to input spatiotemporal sequences and ground truth. The second row indicates sequences predicted by our PredLDM. Following rows are predictions from other models.

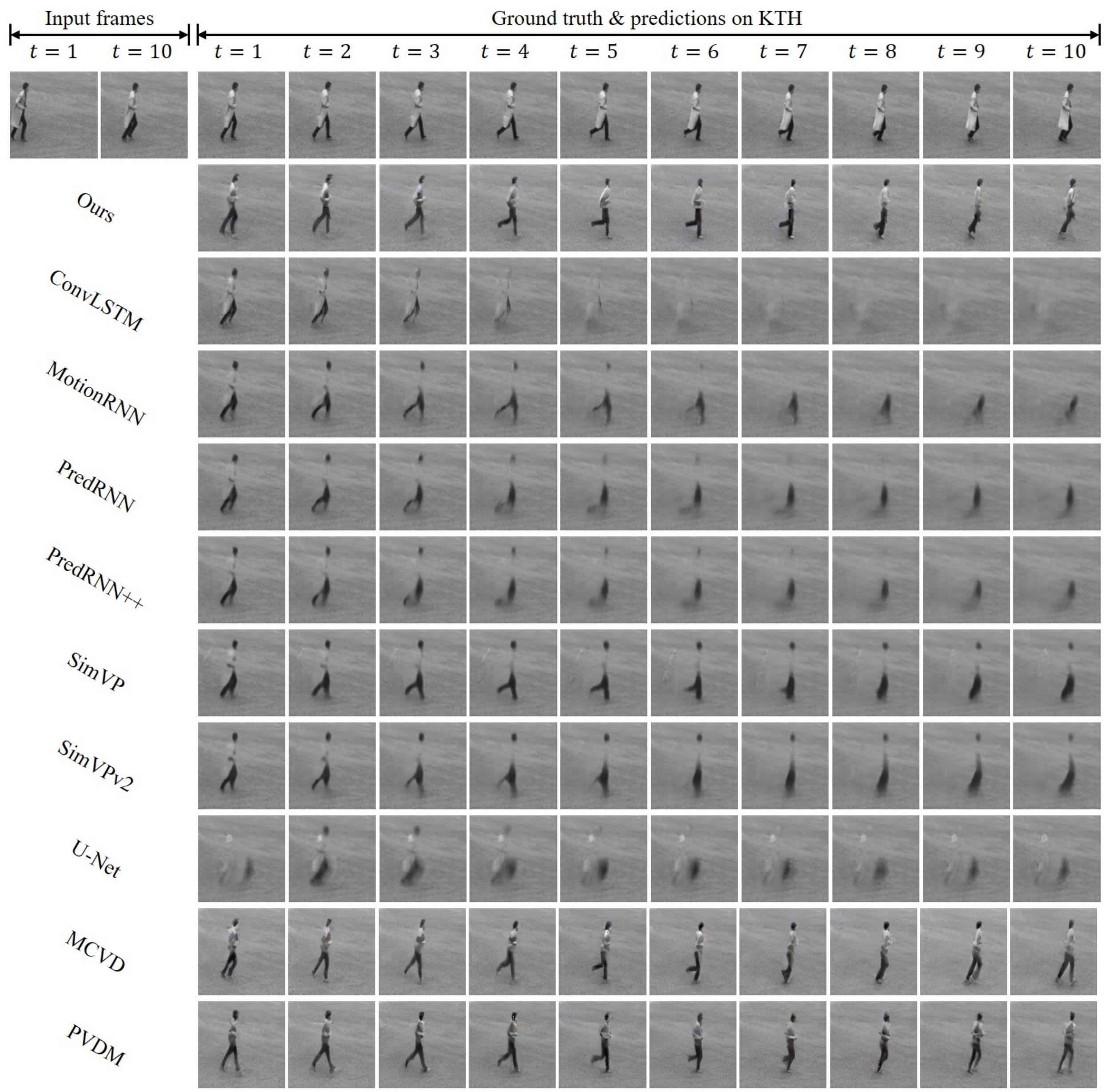

Figure 14: Challenging cases on KTH dataset. The first row refers to input spatiotemporal sequences and ground truth. The second row indicates sequences predicted by our PredLDM. Following rows are predictions from other models.

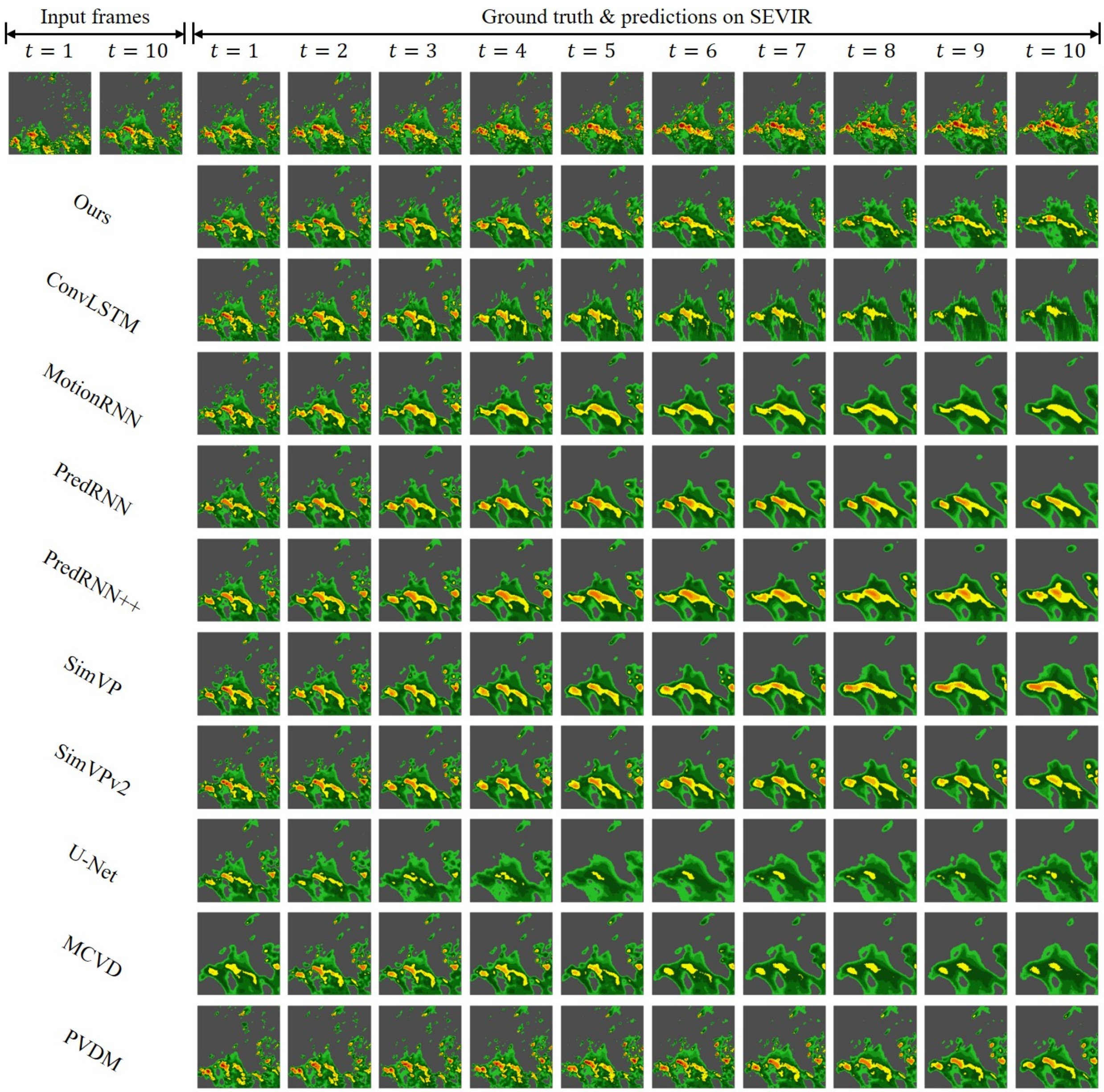

Figure 15: Challenging cases on SEVIR dataset. The first row refers to input spatiotemporal sequences and ground truth. The second row indicates sequences predicted by our PredLDM. Following rows are predictions from other models.

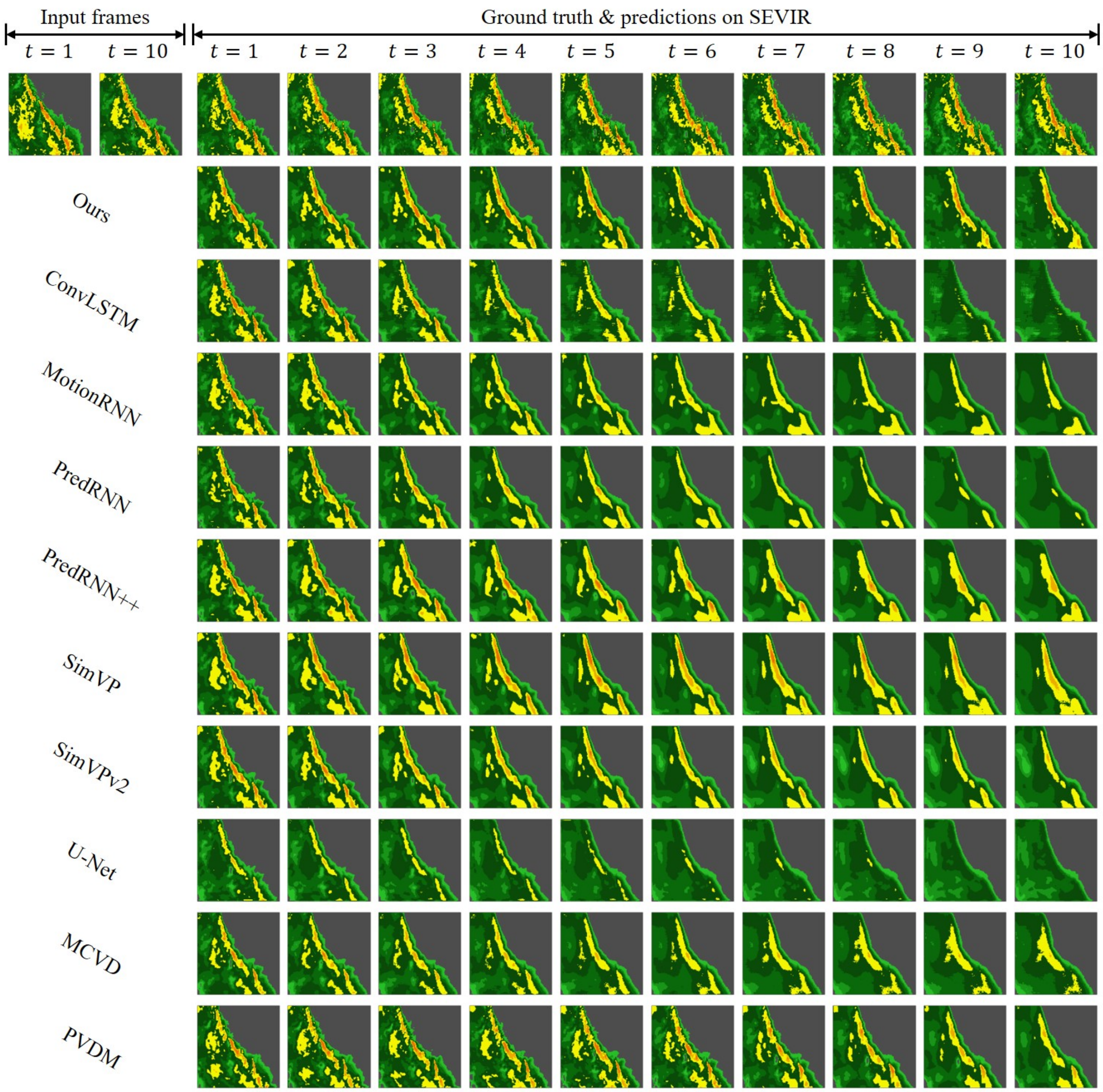

Figure 16: Challenging cases on SEVIR dataset. The first row refers to input spatiotemporal sequences and ground truth. The second row indicates sequences predicted by our PredLDM. Following rows are predictions from other models.