Video Diffusion Models with Local-Global Context Guidance

Siyuan Yang\textsuperscript{1}, Lu Zhang\textsuperscript{2}, Yu Liu\textsuperscript{1*}, Zhizhuo Jiang\textsuperscript{1} and You He\textsuperscript{1}

\textsuperscript{1}Tsinghua University
\textsuperscript{2}Dalian University of Technology

yang-sy21@mails.tsinghua.edu.cn, zhangluu@dlut.edu.cn, \{liuyu77360132, heyou_f\}@126.com, jiangzhizhuo@sz.tsinghua.edu.cn

Abstract

Diffusion models have emerged as a powerful paradigm in video synthesis tasks including prediction, generation, and interpolation. Due to the limitation of the computational budget, existing methods usually implement conditional diffusion models with an autoregressive inference pipeline, in which the future fragment is predicted based on the distribution of adjacent past frames. However, only the conditions from a few previous frames can’t capture the global temporal coherence, leading to inconsistent or even outrageous results in long-term video prediction. In this paper, we propose a Local-Global Context guided Video Diffusion model (LGC-VD) to capture multi-perception conditions for producing high-quality videos in both conditional/unconditional settings. In LGC-VD, the UNet is implemented with stacked residual blocks with self-attention units, avoiding the undesirable computational cost in 3D Conv. We construct a local-global context guidance strategy to capture the multi-perceptual embedding of the past fragment to boost the consistency of future prediction. Furthermore, we propose a two-stage training strategy to alleviate the effect of noisy frames for more stable predictions. Our experiments demonstrate that the proposed method achieves favorable performance on video prediction, interpolation, and unconditional video generation. We release code at https://github.com/exisas/LGC-VD.

1 Introduction

Video prediction aims to generate a set of future frames that are visually consistent with the past content. By capturing the future perceive in a dynamic scene, video prediction has shown great value in many applications such as automaton driving [Hu \textit{et al.}, 2020] and human-machine interaction [Wu \textit{et al.}, 2021]. To make the generated videos more aesthetic, most of the early deep learning methods rely on 3D convolution [Tran \textit{et al.}, 2015] or RNNs [Babaeizadeh \textit{et al.}, 2018; Denton and Fergus, 2018a] to model the temporal coherence of historical frames. However, due to limited long-term memory capacity in both 3D Conv and RNNs, these methods fail to capture the global correspondence of past frames, leading to the inconsistency of both semantics and motion in future frame prediction.

Inspired by the success of Generative Adversarial Networks (GANs) in image synthesis [Vondrick \textit{et al.}, 2016], more recent attempts [Muñoz \textit{et al.}, 2021; Yu \textit{et al.}, 2022] have been made to develop video prediction frameworks on GANs by incorporating Spatio-temporal coherent designs. Thanks to the impressive generative ability of GANs, these methods show great effect in producing videos with natural and consistent content and hold the state-of-the-art in the literature. However, they still suffer from the inherent limitations of GANs in training stability and sample diversity [Dhariwal and Nichol, 2021], which incur unrealistic texture and artifacts in the generated videos.

Recently, Diffusion Probabilistic Models (DPMs) have gained increasing attention and emerged as a new state-of-the-art in video synthesis [Voleti \textit{et al.}, 2022; Ho \textit{et al.}, 2022b] and beyond [Lugmayr \textit{et al.}, 2022; Rombach \textit{et al.}, 2022]. Being likelihood-based models, diffusion models show strengths over GANs in training stability, scalability, and distribution diversity. The core insight is to transform the noise-to-image translation into a progressive denoising process, that is the predicted noise by a parameterized network should be approximate to the predefined distribution in the forward phase. For video prediction, an early attempt by [Ho \textit{et al.}, 2022b] is to extend the standard image dif-
fusion model [Ho et al., 2020] with 3D UNet. The proposed Video Diffusion Model (VDM) thus can denoise the 3D input and return a cleaner video clip iteratively. Despite the impressive results, the sequential denoising mode and the 3D convolution are costly and largely limit the length and resolution of the predicted videos. To produce high-fidelity results, some following works [Ho et al., 2022a; Singer et al., 2022] incorporate extra diffusion models for super-resolution, which further burden the computational cost during training and inference.

To alleviate this issue, MCVD [Voleti et al., 2022] uses stacked residual blocks with self-attention to form the UNet, so that the whole model can be trained and evaluated by limited computation resources within proper inference time. Besides, they propose an autoregressive inference framework where the masked context is taken as a condition to guide the generation of the next clip. As a result, MCVD can solve video prediction, interpolation, and unconditional generation with one unified framework. However, only a few adjacent frames (e.g., two past frames) are fed to the conditional diffusion model, without a global comprehension of the previous fragment. The model might be affected by the noisy past frames and produces inconsistent or even outrageous predictions in long-term videos.

In this paper, we propose a Local-Global Context guided Video Diffusion model (LGC-VD) to capture comprehensive conditions for high-quality video synthesis. Our LGC-VD follows the conditional autoregressive framework of MCVD [Voleti et al., 2022], so the related tasks like video prediction, interpolation, and unconditional generation can be uniformly solved with one framework. As shown in Figure 1, we propose a local-global context guidance strategy, where both global context and local context from previous predictions are incorporated to enhance the consistency of the next video clip. Specifically, the local context from the past fragment is fed to the UNet for full interaction with the prediction during the iterative denoising process. A sequential encoder is used to capture the global embedding of the last fragment, which is integrated with the UNet via latent cross-attention. Furthermore, we build a two-stage training algorithm to alleviate the model’s sensitivity to noisy predictions.

Our main contributions are summarized as follows:

1. A local-global context guidance strategy to capture the multi-perception conditions effectively for more consistent video generation.
2. A two-stage training algorithm, which treats prediction errors as a form of data augmentation, helps the model learn to combat prediction errors and significantly enhance the stability for long video prediction.
3. Experimental results on two datasets demonstrate that the proposed method achieves promising results on video prediction, interpolation, and unconditional generation.

2 Related Work

Generative Models for Video Synthesis. Video synthesis is the task of predicting a sequence of visually consistent frames under the condition of text description, prior frames, or Gaussian noises. To remedy this problem, early methods [Babaeizadeh et al., 2018; Denton and Fergus, 2018a; Castrejón et al., 2019b] build the image-autoregressive model on Recurrent Neural Networks (RNNs) to model the temporal correspondence implicitly. For example, a stochastic video prediction model is proposed in [Babaeizadeh et al., 2018], where the temporal memories embedded in the latent value are used to guide the generation of future frames via a recurrent architecture. Later, Franceschi et al. [2020] propose a stochastic dynamic model by building state-space models in a residual framework. Akan et al. [2021] propose to use LSTM to model the temporal memory in a latent space for frame prediction and optical flow estimation, respectively. Although impressive breakthrough has been made by the aforementioned methods, the recurrent architectures they used usually bring more computational burdens and show limited capacity for modeling the long-term embedding. To alleviate this issue, some recent methods [Yan et al., 2021; Moing et al., 2021; Rakhimov et al., 2021; Weissborn et al., 2019] use Transformer to capture the global space-time correspondence for more accurate video prediction. To further improve the details of the generated videos, some attempts [Vondrick et al., 2016; Luc et al., 2020b] have been made by building video prediction models on Generative Adversarial Networks (GANs). For example, [Tulyakov et al., 2018] propose MoCoGAN to learn to disentangle motion from content in an unsupervised manner. [Muñoz et al., 2021] propose TS-GAN to model spatiotemporal consistency among adjacent frames by building a temporal shift module on a generative network. More recently, DIGAN [Yu et al., 2022] builds an implicit neural representation based on GAN to improve the motion dynamics by manipulating the space and time coordinates.

Diffusion Probabilistic Models. Recently, Diffusion Probabilistic Models (DPMs) [Ho et al., 2020] have received increasing attention due to their impressive ability in image generation. It has broken the long-term domination of GANs and become the new state-of-the-art protocol in many computer vision tasks, such as image/video/3D synthesis [Dhariwal and Nichol, 2021; Saharia et al., 2022; Ramesh et al., 2022; Nichol et al., 2021], image inpainting [Lugmayr et al., 2022] and super-resolution [Rombach et al., 2022]. Based on the predefined step-wise noise in the forward process, DPMs [Ho et al., 2020; Song et al., 2020a; Nichol and Dhariwal, 2021; Song et al., 2020b] leverage a parametric U-Net to denoise the input Gaussian distribution iteratively in the reverse process. For unconditional video generation, [Ho et al., 2022b] extends the original UNet to 3D format [Čiçek et al., 2016] to process video data. For conditional video prediction tasks, a naive approach is to use the unconditional video generation model directly by means of conditional sampling as RePaint [Lugmayr et al., 2022]. However, this approach relies on a well-trained unconditional generative model, which is computationally resource-intensive to obtain. To overcome this problem, [Čiçek et al., 2016] and [Harvey et al., 2022] propose a diffusion-based architecture for video prediction and infilling. For a video with
Figure 2: Left: In the first stage, the global sequence encoder extracts global context embeddings $z^{i-1}$ from a fixed tensor. Conditional frames are selected from the ground truth, and after concatenation with positional masks, local context $y'_{m-1}$ is obtained. Right: In the second stage, the sequence encoder extracts global context embeddings from the entire predicted $x^0_{m-1}$, and the conditional frames are selected from the predicted $x^i_{m-1}$. In the test phase, our model performs iterative denoising on Gaussian noise to obtain the prediction of the current fragment $x_0$, after which it moves on to the following fragment. In the training phase, our model processes the noisy current fragment $x_t$ and predicts the noise $\epsilon_t$ to obtain $x_0$ directly by Eq.7.

Given a sample from data distribution $x_0 \sim q(x)$, the forward process of the diffusion model destroys the structure in data by adding Gaussian noise to $x_0$ iteratively. The Gaussian noise is organized according to a variance schedule $\beta_1, ..., \beta_T$, resulting in a sequence of noisy samples $x_1, ..., x_T$, where $T$ is diffusion step. When $T \to \infty$, $x_T$ is equivalent to an isotropic Gaussian distribution. This forward process can be defined as

$$ q \left( x_{1:T} \right) := \prod_{t=1}^{T} q \left( x_t \mid x_{t-1} \right) $$  \hspace{1cm} (1)$$

$$ q \left( x_t \mid x_{t-1} \right) := \mathcal{N} \left( x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I \right) $$  \hspace{1cm} (2)$$

With the transition kernel above, we can sample $x_t$ at any time step $t$:

$$ q_t \left( x_t \mid x_0 \right) := \mathcal{N} \left( x_t; \sqrt{\alpha_t} x_0, (1 - \alpha_t) I \right) $$  \hspace{1cm} (3)$$

where $\alpha_t = \prod_{s=1}^{t} (1 - \beta_s)$. In the training phase, the reverse process tries to trace back from the isotropic Gaussian noise $x_T \sim \mathcal{N} \left( x_T; 0, I \right)$ to the initial sample $x_0$. Since the exact reverse distribution $q \left( x_{t-1} \mid x_t \right)$ cannot be obtained, we use a Markov chain with learned Gaussian transitions to replace it:

$$ p_\theta \left( x_{0:T} \right) := \prod_{t=1}^{T} p_\theta \left( x_{t-1} \mid x_t \right) $$  \hspace{1cm} (4)$$

$$ p_\theta \left( x_{t-1} \mid x_t \right) := \mathcal{N} \left( x_{t-1}; \mathbf{\mu}_\theta \left( x_t, x_0 \right), \Sigma_\theta \left( x_t, t \right) \right) $$  \hspace{1cm} (5)$$

In practice, we use KL divergence to get $p_\theta \left( x_{t-1} \mid x_t \right)$ to estimate the corresponding forward process posteriors:

$$ q \left( x_{t-1} \mid x_t, x_0 \right) := \mathcal{N} \left( x_{t-1}; \tilde{\mathbf{\mu}} \left( x_t, x_0 \right), \beta_I \right) $$  \hspace{1cm} (6)$$

where $\tilde{\mathbf{\mu}} \left( x_t, x_0 \right) := \frac{\sqrt{\alpha_t} \beta_t}{1 - \alpha_t} x_0 + \frac{\sqrt{\alpha_t} (1 - \beta_t)}{1 - \alpha_t} x_t$, $\alpha_t = 1 - \beta_t$ and $\beta_t := \frac{1 - \alpha_t}{1 - \alpha_{t+1}} \beta_{t+1}$. Since $x_0$ given $x_T$ is unknown in the
sampling phase, We use a time-conditional neural network parameterized by $\theta$ to estimate the noise $\epsilon_t$, then we can get

$$\hat{x}_0 = \frac{1}{\sqrt{\alpha_t}} (x_t - \sqrt{1 - \alpha_t} \epsilon_t)$$

(7)

by Eq.3. The loss can be formulated as

$$L(\theta) = E_{x_0, \epsilon} \left[ \left\| \epsilon - \epsilon_\theta \left( \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon, t \right) \right\|^2 \right]$$

(8)

Eventually, we can reverse the diffusion process by replacing $x_0$ in Eq.6 with $\hat{x}_0$ to reconstruct the structure in data from Gaussian noise.

$$y^{i-1} \cdot m^{i-1}, x^{i-1}_0, x^{i-1}_0, y^i \cdot m^i, x^i_0, \epsilon_{i1}, \epsilon_{i2}$$

### 3.2 LGC-VD for Video Synthesis Tasks

In contrast to [Voleti et al., 2022], our model extracts both local and global context from the past fragment, to boost the consistency of future prediction. For video prediction task, given $p$ start frames, our model aims to produce a fragment with $k$ frames each time and finally propose a $(n = k \times m + p)$-length video, where $m$ is the number of fragments.

As shown in Figure 2, to predict the fragment $x^i$ at time $t \in \{1, ..., m - 1\}$, we use the local context from last fragment $x^{i-1}$ as local condition, which is represented as $y^i$. Besides, we employ a sequential encoder to extract the global feature from the last fragment as a global condition, which is represented as $z^i$. Since no past fragment is given at $i = 0$, the sequential encoder takes only the input as a fixed tensor $U$ that shares the same shape as $x^i$. The local condition and the global embedding are incorporated in the UNet to guide the generation of the future fragment.

We introduce a positional mask on local conditional frames for the more flexible switch to different tasks. Specifically, each selected conditional frame is concatenated with a positional mask $m \in \mathbb{R}^{1 \times H \times W}$, whose pixels are filled with a fixed value of $\frac{1}{m+1}, j \in [-1, k]$. Then, our model can take any frame from the fragment as a condition and flexibly conduct future prediction or filling. Besides, for unconditional generation, we set $j = -1$ to indicate a fixed tensor $U$ as initial conditions. Finally, the Eq. 8 can be rewritten as

$$L(\theta) = E_{x^i, \epsilon} \left[ \left\| \epsilon - \epsilon_\theta \left( \sqrt{\alpha_t} x^i + \sqrt{1 - \alpha_t} \epsilon, t \right) \right\|^2 \right]$$

(9)

Where $y^i \cdot m^i$ is the concatenation of $y^i$ and $m^i$. In practice, the local condition $y^i \cdot m^i$ is incorporated into the network via channel-wise concatenation, whereas the global condition $z^i$ is mapped to the intermediate layers of the UNet via a cross attention layer:

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QKT}{\sqrt{d}} \right) \cdot V$$

(10)

$$Q = W_Q^{(i)} \cdot f^i, K = W_K^{(i)} \cdot z^i, V = W_V^{(i)} \cdot z^i$$

(11)

where $z^i$ is the feature of $x^{i-1}$ by sequential encoder. $f^i$ is the feature of current fragment by UNet encoder and $W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}$ are learnable projection matrices.

### 3.3 Two-stage Training

In previous autoregressive frameworks [Voleti et al., 2022; Ho et al., 2022b; Höppe et al., 2022], the ground truths of previous frames are directly used as conditions during training. However, undesirable errors in content or motion would accumulate and affect long-term predictions. To alleviate this issue, we propose a two-stage training strategy.

In the first stage, the conditional frames $y^0$ are randomly selected from ground truth $x^0$, and the sequential encoder takes the input as a constant tensor $U$. The model then outputs the first predicted video sequence $\hat{x}^0$. In the second stage, the conditional frames $y^i$ are the last few frames from the previous prediction $\hat{x}^0$, the conditional feature $z^0$ is the result of encoding $x^0$ by the sequence encoder. With this setting, the prediction errors in the training phase are included and are treated as a form of data augmentation to improve the network’s ability for long video prediction. Therefore, the first stage of training endows our model with the flexibility to make future predictions or infillings, and the second stage enhances the model’s ability to predict long videos.

Since iteratively denoising in the training phase will greatly increase the computational burden, we use Eq.7 to directly obtain the prediction of the current fragment $\hat{x}_0$ from the noisy fragment $x_t$. A natural question is whether using the $x_0$ obtained through Eq.7 instead of the $x_0$ obtained by iterative denoising can help the model learn to combat prediction error. To evaluate this, we randomly select a training sample and display the results obtained by Eq.7 (top row) and the results obtained by iterative denoising (middle row) in Figure 3. As can be seen, the results of both predictions show blurry effects in uncertainty areas (e.g., the occluded region of the background or the regions with object interaction). Therefore, this substitution can be considered a legitimate data enhancement.

### 4 Experiments

We present the experimental results of video prediction on Cityscapes [Cordts et al., 2016], and the experimental results
of video prediction, video generation, video interpolation on BAIR [Ebert et al., 2017].

4.1 Datasets and Metrics

Dataset. Cityscapes [Cordts et al., 2016] is a large-scale dataset that contains a diverse set of stereo video sequences recorded in street scenes from 50 different cities. We use the same data preprocessing method as [Yang et al., 2022; Voleti et al., 2022], from the official website, we obtain the leftImg8bit_sequence_trainvaltest. This package includes a training set of 2975 videos, a validation set of 500 videos, and a test set of 1525 videos, each with 30 frames. All videos are center cropped and downsampled to 128×128. BAIR Robot Pushing [Ebert et al., 2017] is a common benchmark in the video literature, which consists of roughly 44000 movies of robot pushing motions at 64×64 spatial resolution.

Metrics. In order to compare our LGC-VD with prior works, we measure our experimental results in PSNR, LPIPS, SSIM, and FVD [Unterthiner et al., 2018]. FVD compares statistics in the latent space of an Inflated 3D ConvNet (I3D) trained on Kinetics-400, which measures time coherence and visual quality.

4.2 Implementation Details

We use a U-Net that is composed of encoding and decoding for upsample and downsample, respectively. Each part contains two residual blocks and two attention blocks. The middle layer of UNet is composed of one residual block followed by an attention block. The sequential encoder uses the front half and middle layer of diffusion UNet, with a total of three residual blocks and three attention modules. Since our sequential encoder only needs to be called once when iteratively denoising, the added time in the test phase is negligible. Besides, in our experiments, we found that when ε — prediction is coupled with our model’s spatial-temporal attention block, particularly when trained at resolutions of 128 × 128 and higher, the predicted video has brightness variations and occasionally color inconsistencies between frames. We use v-prediction [Salimans and Ho, 2022] to overcome this problem.

All of our models are trained with Adam on 4 NVIDIA Tesla V100s with a learning rate of 1e-4 and a batch size of 32 for Cityscapes and 192 for BAIR. We use the cosine noise schedule in the training phase and set the diffusion step T to 1000. For both datasets, we set the total video length L to 14, the video length N for each stage to 8, and the number of conditional frames K to 2. At testing, we sample 100 steps using DDPM.

4.3 Evaluation on Video Prediction

Table 1 lists all the metric scores of our model and other competitors on Cityscapes. For video prediction, our model significantly outperforms MCVD [Voleti et al., 2022] in FVD, LPIPS, and SSIM. Despite the fact that our SSIM is slightly lower than GHVAE [Wu et al., 2021], it significantly outperforms GHVAE in FVD and LPIPS.

Table 3 lists all the metric scores of our model and other methods on BAIR. To compare our model with previous work on video prediction, we evaluated our model under two settings of start frames. With the first two frames serving as references, the two evaluation methods predict 14 frames and 28 frames, respectively. As can be seen, our model is slightly lower than MCVD in terms of FVD, but much better than MCVD in terms of PSNR and SSIM. Besides, Table 3 also shows that our model’s result under just one start frame (p = 1), which outperforms other methods by large margin.

The perceptual performance is also verified visually in Figure 4. As we can see, our LGC-VD is effective at producing high-quality videos with clear details, while MCVD tends to produce blurry results on the regions with potentially fast motion. Besides, our LGC-VD alleviates the issue of inconsistent brightness in MCVD.

4.4 Evaluation on Video Interpolation

The experimental results of our model’s video interpolation on BAIR are shown in Table 2. Compared with MCVD, our model requires fewer reference images, predicts more intermediate frames, and outperforms MCVD in terms of SSIM and PSNR. Even when compared to architectures designed for video interpolation [Xu et al., 2020; Niklaus et al., 2017], our model obtains SOTA results for both metrics, 26.732 and 0.952, respectively.

4.5 Evaluation on Video Generation

We also present the results of unconditional video generation on BAIR. In this setting, no conditional frames are given and

<table>
<thead>
<tr>
<th>Cityscapes</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>LGC-VD (Ours)</td>
</tr>
<tr>
<td>SVG-LP [Denton and Fergus, 2018b]</td>
</tr>
<tr>
<td>vRNN IL [Castrejon et al., 2019a]</td>
</tr>
<tr>
<td>Hier-vRNN [Castrejon et al., 2019a]</td>
</tr>
<tr>
<td>GHVAE [Wu et al., 2021]</td>
</tr>
<tr>
<td>MCVD spatin [Voleti et al., 2022]</td>
</tr>
<tr>
<td>MCVD concat [Voleti et al., 2022]</td>
</tr>
<tr>
<td>LGC-VD (Ours)</td>
</tr>
</tbody>
</table>

Table 1: Video prediction on Cityscapes. Predicting k frames using the first p frames as a condition, then recursively predicting n frames. We illustrate video prediction results under various lengths of conditions by p = 1 and p = 2.
Figure 4: Qualitative results of video prediction on Cityscapes and BAIR. From top to bottom: Ground Truth (Top Row), MCVD spatin (Second Row), MCVD concat (Third Row), Our Method (Bottom Row). On BAIR, we observe that the MCVD prediction becomes blurry over time while our LGC-VD predictions remain crisp. On Cityscapes, the prediction of MCVD shows significant brightness biases, while our LGC-VD performs better in both motion and content consistency.

<table>
<thead>
<tr>
<th>BAIR</th>
<th>(p + f/k/n)</th>
<th>PSNR↑ / SSIM↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVG-LP</td>
<td>18 / 7 / 100</td>
<td>18.648 / 0.846</td>
</tr>
<tr>
<td>SepConv</td>
<td>18 / 7 / 100</td>
<td>21.615 / 0.877</td>
</tr>
<tr>
<td>SDVI full</td>
<td>18 / 7 / 100</td>
<td>21.432 / 0.880</td>
</tr>
<tr>
<td>SDVI</td>
<td>18 / 7 / 100</td>
<td>19.694 / 0.852</td>
</tr>
<tr>
<td>MCVD</td>
<td>4 / 5 / 100</td>
<td>25.162 / 0.932</td>
</tr>
<tr>
<td><strong>LGC-VD (ours)</strong></td>
<td>2 / 6 / 100</td>
<td><strong>26.732/0.952</strong></td>
</tr>
</tbody>
</table>

Table 2: Video interpolation on BAIR. Given \(p\) past with \(f\) future frames, interpolating \(k\) frames. We report average of the best metrics out of \(n\) trajectories per test sample.

The models need to synthesize the videos from only Gaussian noise. As shown in Table 4, our model first generates 8-frame video sequences in an unconditional manner, then generates 6-frame video sequences each time using our autoregressive inference framework, and finally creates video sequences with 30 frames. Our model yields a FVD of 250.71, which is significantly improved compared to the 348.2 of MCVD.

4.6 Ablation Study

The proposed LGC-VD is built upon an autoregressive inference framework, where several past frames are taken as a condition to boost the generation of the next video fragment. In LGC-VD, we propose a hierarchical context guidance strategy to incorporate the local context from a few past frames with the global context of historical clips. Besides, considering that the prediction would be affected by the noisy conditions, a two-stage training algorithm is proposed to facilitate the robustness of the model for more stable prediction. To demonstrate the efficacy of the two contributions, we conduct ablation studies on Cityscapes. As shown in Table 5, we implement a baseline by removing the sequential encoder and the two-stage training strategies. Specifically, the UNet takes only the past \(p\) frames as a condition to model the local guidance, and the second training is skipped. From the comparison among “+two-stage training”, “+sequential
In this paper, we propose local-global context guidance to comprehensively construct the multi-perception embedding from the previous fragment, for high-quality video synthesis. We also propose a two-stage training strategy to alleviate the effect of noisy conditions and boost the model to produce more stable predictions. Our experiments demonstrate that the proposed method achieves state-of-the-art performance on video prediction, as well as favorable performance on interpolation and unconditional video generation.

To summarize, our method makes a further improvement in performance on video prediction, as well as favorable performance on interpolation and unconditional video generation.

**Acknowledgements**

This study was co-supported by the National Key R&D Program of China under Grant 2021YFA0715202, the National Natural Science Foundation of China (Nos. 62293544, 62022092), as well as partially supported by the National Postdoctoral Program of China for Innovative talent (BX2021051) and the National Nature Science Foundation of China (62206039).
References


