

Video Diffusion Models with Local-Global Context Guidance

Siyuan Yang¹, Lu Zhang², Yu Liu^{1*}, Zhizhuo Jiang¹ and You He¹

¹Tsinghua University

²Dalian 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 *et al.*, 2020] and human-machine interaction [Wu *et al.*, 2021]. To make the generated videos more aesthetic, most of the early deep learning methods rely on 3D convolution [Tran *et al.*, 2015] or RNNs [Babaeizadeh *et al.*, 2018; Denton and Fergus, 2018a] to model the temporal coherence

*Corresponding authors

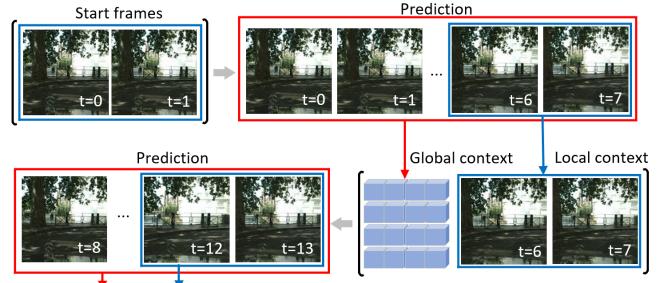


Figure 1: We propose an autoregressive inference framework, where both global context and local context from previous predictions are incorporated to enhance the consistency of the next video clip.

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 *et al.*, 2016], more recent attempts [Muñoz *et al.*, 2021; Yu *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 *et al.*, 2022; Ho *et al.*, 2022b] and beyond [Lugmayr *et al.*, 2022; Rombach *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 *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; Weissenborn *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 [Çiçek *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, [Çiçek *et al.*, 2016] and [Harvey *et al.*, 2022] propose a diffusion-based architecture for video prediction and infilling. For a video with

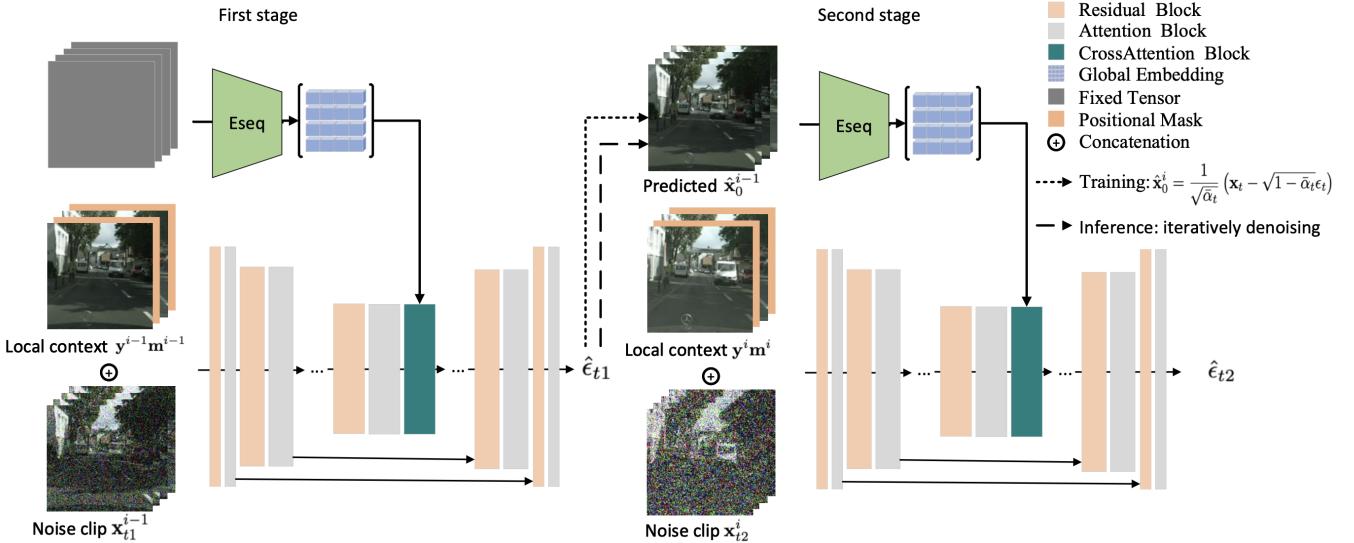


Figure 2: Left: In the first stage, the global sequence encoder extracts global context embeddings \mathbf{z}^{i-1} from a fixed tensor. Conditional frames are selected from the ground truth, and after concatenation with positional masks, local context $\mathbf{y}^{i-1}\mathbf{m}^{i-1}$ is obtained. Right: In the second stage, the sequence encoder extracts global context embeddings from the entire predicted $\hat{\mathbf{x}}_0^{i-1}$, and the conditional frames are selected from the predicted $\hat{\mathbf{x}}_0^{i-1}$. In the test phase, our model performs iterative denoising on Gaussian noise to obtain the prediction of the current fragment \mathbf{x}_0 , after which it moves on to the following fragment. In the training phase, our model processes the noisy current fragment \mathbf{x}_t and predicts the noise ϵ_t to obtain \mathbf{x}_0 directly by Eq.7.

m frames, they randomly select n frames as the conditional frames, which are kept unchanged in the forward and backward processes, and carry out diffusion and denoising on the remaining $m - n$ frames. MCVD proposed by [Voleti *et al.*, 2022] concatenates all video frames in channel dimension, and represents the video as four-dimensional data. This work uses 2D Conv instead of 3D Conv, which greatly reduces the computational burden without compromising the quality of the results, and achieves SOTA results on video prediction and interpolation.

3 Method

In this paper, we propose a Local-Global Context guided Video Diffusion model (LGC-VD) to address video synthesis tasks, including video prediction, interpolation, and unconditional generation. Our model is built upon an autoregressive inference framework, where the frames from the previous prediction are used as conditions to guide the generation of the next clip. We construct a local-global context guidance strategy to achieve comprehensive embeddings of the past fragment to boost the consistency of future prediction. Besides, we propose a two-stage training algorithm to alleviate the influence of noisy conditional frames and facilitate the robustness of the diffusion model for more stable prediction. Below, we first review the diffusion probabilistic models.

3.1 Background

Given a sample from data distribution $\mathbf{x}_0 \sim q(\mathbf{x})$, the forward process of the diffusion model destroys the structure in data by adding Gaussian noise to \mathbf{x}_0 iteratively. The Gaussian noise is organized according to a variance schedule

β_1, \dots, β_T , resulting in a sequence of noisy samples $\mathbf{x}_1, \dots, \mathbf{x}_T$, where T is diffusion step. When $T \rightarrow \infty$, \mathbf{x}_T is equivalent to an isotropic Gaussian distribution. This forward process can be defined as

$$q(\mathbf{x}_{1:T}) := \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}) \quad (1)$$

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \quad (2)$$

With the transition kernel above, we can sample \mathbf{x}_t at any time step t :

$$q_t(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}) \quad (3)$$

where $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$. In the training phase, the reverse process tries to trace back from the isotropic Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ to the initial sample \mathbf{x}_0 . Since the exact reverse distribution $q(\mathbf{x}_{t-1} | \mathbf{x}_t)$ cannot be obtained, we use a Markov chain with learned Gaussian transitions to replace it:

$$p_\theta(\mathbf{x}_{0:T}) := \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) \quad (4)$$

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (5)$$

In practice, we use KL divergence to get $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$ to estimate the corresponding forward process posteriors :

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}) \quad (6)$$

where $\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1 - \bar{\alpha}_t} \mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{x}_t$, $\alpha_t = 1 - \beta_t$ and $\tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$. Since \mathbf{x}_0 given \mathbf{x}_T is unknown in the

sampling phase, We use a time-conditional neural network parameterized by θ to estimate the noise ϵ_t , then we can get

$$\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} (\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_t) \quad (7)$$

by Eq.3. The loss can be formulated as

$$L(\theta) = \mathbf{E}_{\mathbf{x}_0, \epsilon} \left[\|\epsilon - \epsilon_\theta (\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|^2 \right] \quad (8)$$

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

$$\mathbf{y}^{i-1} \mathbf{m}^{i-1}, \mathbf{x}_{t1}^{i-1}, \hat{\mathbf{x}}_0^{i-1}, \mathbf{y}^i \mathbf{m}^i, \mathbf{x}_{t2}^i, \hat{\epsilon}_{t1}, \hat{\epsilon}_{t2}$$

3.2 LGC-VD for Video Synthesis Tasks

In contrast to [Voeleti *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 \mathbf{x}^i at time $i \in \{0, \dots, m-1\}$, we use the local context from last fragment \mathbf{x}^{i-1} as local condition, which is represented as \mathbf{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 \mathbf{z}^i . Since no past fragment is given at $i=0$, the sequential encoder takes only the input as a fixed tensor \mathbf{U} that shares the same shape as \mathbf{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 $\mathbf{m} \in \mathbf{R}^{1 \times H \times W}$, whose pixels are filled with a fixed value of $\frac{j+1}{k+1}, j \in [-1, k]$. Then, our model can take any frame from the fragment as a condition and flexibly conduct future prediction or infilling. Besides, for unconditional generation, we set $j = -1$ to indicate a fixed tensor \mathbf{U} as initial conditions. Finally, the Eq. 8 can be rewritten as

$$L(\theta) = \mathbf{E}_{\mathbf{x}^i, \epsilon} \left[\|\epsilon - \epsilon_\theta (\sqrt{\bar{\alpha}_t} \mathbf{x}^i + \sqrt{1 - \bar{\alpha}_t} \epsilon, t, \mathbf{y}^i \mathbf{m}^i, \mathbf{z}^i)\|^2 \right] \quad (9)$$

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

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d}} \right) \cdot V \quad (10)$$

$$Q = W_Q^{(i)} \cdot \mathbf{f}^i, K = W_K^{(i)} \cdot \mathbf{z}^i, V = W_V^{(i)} \cdot \mathbf{z}^i \quad (11)$$

where \mathbf{z}^i is the feature of \mathbf{x}^{i-1} by sequential encoder. \mathbf{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.

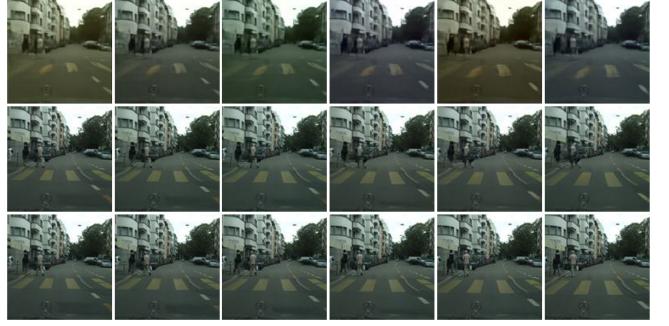


Figure 3: Top row shows the output of the first stage, which is referring to the start frames, adding noise to current fragment \mathbf{x}_0 using Eq.3 to obtain noisy fragment \mathbf{x}_{592} , then predicting \mathbf{x}_0 directly using Eq.7. The middle row shows the prediction of \mathbf{x}_0 obtained by iterative denoising from Gaussian noise using the start frames as a reference. The third row shows the ground truth. The six columns are frames 2-7.

3.3 Two-stage Training

In previous autoregressive frameworks [Voeleti *et al.*, 2022; Ho *et al.*, 2022b; Höpke *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 \mathbf{y}^0 are randomly selected from ground truth \mathbf{x}^0 , and the sequential encoder takes the input as the constant tensor \mathbf{U} . The model then outputs the first predicted video sequence $\hat{\mathbf{x}}^0$. In the second stage, the conditional frames \mathbf{y}^1 are the last few frames from the previous prediction $\hat{\mathbf{x}}^0$, the conditional feature \mathbf{z}^0 is the result of encoding $\hat{\mathbf{x}}^0$ 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 $\hat{\mathbf{x}}_0$ from the noisy fragment \mathbf{x}_t . A natural question is whether using the $\hat{\mathbf{x}}_0$ obtained through Eq.7 instead of the $\hat{\mathbf{x}}_0$ 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

Cityscapes	p	k	n	FVD \downarrow	LPIPS \downarrow	SSIM \uparrow
LGC-VD (Ours)	1	7	28	182.8	0.084 ± 0.03	0.693 ± 0.08
SVG-LP [Denton and Fergus, 2018b]	2	10	28	1300.26	0.549 ± 0.06	0.574 ± 0.08
vRNN 1L [Castrejon <i>et al.</i> , 2019a]	2	10	28	682.08	0.304 ± 0.10	0.609 ± 0.11
Hier-vRNN [Castrejon <i>et al.</i> , 2019a]	2	10	28	567.51	0.264 ± 0.07	0.628 ± 0.10
GHVAE [Wu <i>et al.</i> , 2021]	2	10	28	418.00	0.193 ± 0.014	0.740 ± 0.04
MCVD spatin [Voleti <i>et al.</i> , 2022]	2	5	28	184.81	0.121 ± 0.05	0.720 ± 0.11
MCVD concat [Voleti <i>et al.</i> , 2022]	2	5	28	141.31	0.112 ± 0.05	0.690 ± 0.12
LGC-VD (Ours)	2	6	28	124.62	0.069 ± 0.03	0.732 ± 0.09

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 $p = 2$.

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 128x128. 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 64x64 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 $\epsilon - predicton$ 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

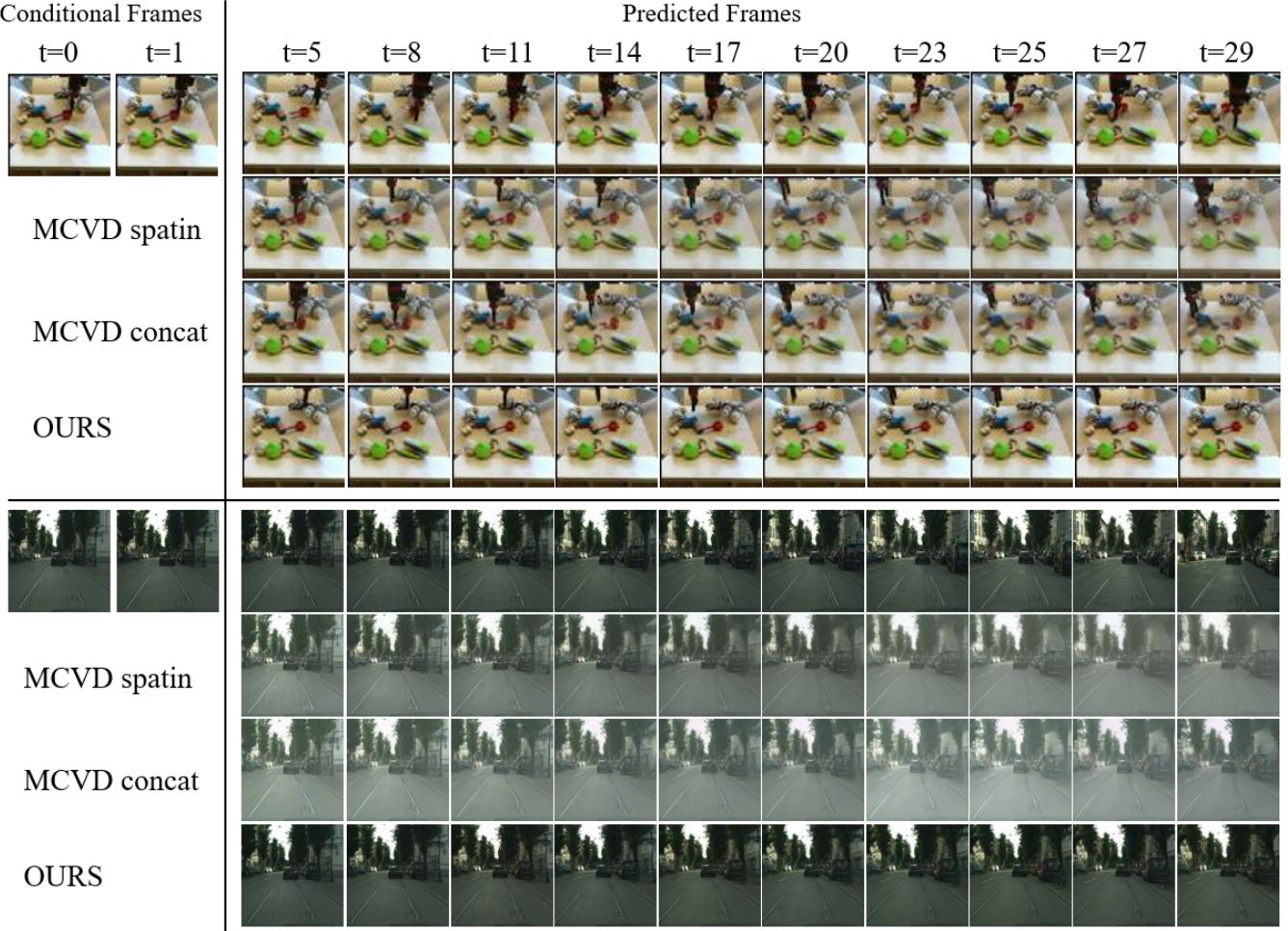


Figure 4: Qualitative results of video prediction on Cityscapes and BAIR. From top to bottom: Ground Truth (Top Row), MCVD spatin (Second Row), MCVD concat (Third Row), Our Method (Bottom Row). On BAIR, we observe that the MCVD prediction becomes blurry over time while our LGC-VD predictions remain crisp. On Cityscapes, the prediction of MCVD shows significant brightness biases, while our LGC-VD performs better in both motion and content consistency.

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

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

BAIR	<i>p</i>	<i>k</i>	<i>n</i>	FVD ↓	PSNR ↑	SSIM ↑
LVT [Rakhimov <i>et al.</i> , 2021]	1	15	15	125.8	—	—
DVD-GAN-FP [Clark <i>et al.</i> , 2019]	1	15	15	109.8	—	—
TrIVD-GAN-FP [Luc <i>et al.</i> , 2020a]	1	15	15	103.3	—	—
VideoGPT [Yan <i>et al.</i> , 2021]	1	15	15	103.3	—	—
CCVS [Moing <i>et al.</i> , 2021]	1	15	15	99.0	—	—
Video Transformer [Weissenborn <i>et al.</i> , 2019]	1	15	15	96	—	—
FitVid [Babaeizadeh <i>et al.</i> , 2021]	1	15	15	93.6	—	—
MCVD spatin past-mask [Voleti <i>et al.</i> , 2022]	1	5	15	96.5	18.8	0.828
MCVD concat past-mask [Voleti <i>et al.</i> , 2022]	1	5	15	95.6	18.8	0.832
MCVD concat past-future-mask [Voleti <i>et al.</i> , 2022]	1	5	15	89.5	16.9	0.780
LGC-VD (Ours)	1	7	15	80.9	21.8	0.891
SAVP [Lee <i>et al.</i> , 2018]	2	14	14	116.4	—	—
MCVD spatin past-mask [Voleti <i>et al.</i> , 2022]	2	5	14	90.5	19.2	0.837
MCVD concat past-future-mask [Voleti <i>et al.</i> , 2022]	2	5	14	89.6	17.1	0.787
MCVD concat past-mask [Voleti <i>et al.</i> , 2022]	2	5	14	87.9	19.1	0.838
LGC-VD (Ours)	2	6	14	76.5	21.9	0.892
SVG-LP [Denton and Fergus, 2018b]	2	10	28	256.6	—	0.816
SLAMP [Akan <i>et al.</i> , 2021]	2	10	28	245.0	19.7	0.818
SAVP [Lee <i>et al.</i> , 2018]	2	10	28	143.4	—	0.795
Hier-vRNN [Castrejon <i>et al.</i> , 2019a]	2	10	28	143.4	—	0.822
MCVD spatin past-mask [Voleti <i>et al.</i> , 2022]	2	5	28	127.9	17.7	0.789
MCVD concat past-mask [Voleti <i>et al.</i> , 2022]	2	5	28	119.0	17.7	0.797
MCVD concat past-future-mask [Voleti <i>et al.</i> , 2022]	2	5	28	118.4	16.2	0.745
LGC-VD (Ours)	2	6	28	120.1	20.39	0.863

Table 3: Video prediction on BAIR. Predicting k frames using the first p frames as a condition, then recursively predicting n frames in total.

BAIR	<i>k</i>	<i>n</i>	FVD ↓
MCVD spatin [Voleti <i>et al.</i> , 2022]	5	30	399.8
MCVD concat [Voleti <i>et al.</i> , 2022]	5	30	348.2
LGC-VD (ours)	6	30	250.7

Table 4: Unconditional video generation on BAIR. Generating k frames every time, then recursively predicting n frames.

encoder” and “baseline”, we can observe that the proposed two-stage training strategy and local-global context guidance can significantly improve the stability of video prediction. By combining the sequential encoder and the local-global context guidance, our model achieves further improvement, especially on FVD, which demonstrates that the global memory from the sequential encoder can collaborate with the original local context to achieve more consistent results.

5 Conclusion and Discussion

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 per-

Cityscapes	<i>p/k/n</i>	FVD / LPIPS / SSIM
Baseline	2/6/28	276.27 / 0.111 / 0.708
+ two-stage training	2/6/28	141.91 / 0.071 / 0.742
+ sequential encoder	2/6/28	153.08 / 0.067 / 0.750
full	2/6/28	124.62 / 0.069 / 0.732

Table 5: Ablation study on Cityscape. Note that we conduct experiments on video prediction, where 2 initial frames ($p = 2$) are conditions to predict 6 frames ($k=6$) and propose a 28-frame video ($n = 28$).

formance on video prediction, as well as favorable performance on interpolation and unconditional video generation. To summarize, our method makes a further improvement in the condition formulation and the training stability for memory-friendly video diffusion methods.

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

- [Akan *et al.*, 2021] Adil Kaan Akan, Erkut Erdem, Aykut Erdem, and Fatma Güney. SLAMP: stochastic latent appearance and motion prediction. In *International Conference on Computer Vision*, pages 14708–14717, 2021.
- [Babaeizadeh *et al.*, 2018] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018.
- [Babaeizadeh *et al.*, 2021] 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.
- [Castrejon *et al.*, 2019a] Lluís Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7608–7617, 2019.
- [Castrejón *et al.*, 2019b] Lluís Castrejón, Nicolas Ballas, and Aaron C. Courville. Improved conditional vrnns for video prediction. In *International Conference on Computer Vision*, pages 7607–7616, 2019.
- [Çiçek *et al.*, 2016] Özgün Çiçek, Ahmed Abdulkadir, Søren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.
- [Clark *et al.*, 2019] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019.
- [Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [Denton and Fergus, 2018a] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In Jennifer G. Dy and Andreas Krause, editors, *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1182–1191, 2018.
- [Denton and Fergus, 2018b] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International conference on machine learning*, pages 1174–1183. PMLR, 2018.
- [Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- [Ebert *et al.*, 2017] Frederik Ebert, Chelsea Finn, Alex X Lee, and Sergey Levine. Self-supervised visual planning with temporal skip connections. In *CoRL*, pages 344–356, 2017.
- [Franceschi *et al.*, 2020] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3233–3246, 2020.
- [Harvey *et al.*, 2022] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022.
- [Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [Ho *et al.*, 2022a] 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*, 2022.
- [Ho *et al.*, 2022b] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.
- [Höppe *et al.*, 2022] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv preprint arXiv:2206.07696*, 2022.
- [Hu *et al.*, 2020] Anthony Hu, Fergal Cotter, Nikhil Mohan, Corina Gurau, and Alex Kendall. Probabilistic future prediction for video scene understanding. In *European Conference on Computer Vision*, pages 767–785. Springer, 2020.
- [Lee *et al.*, 2018] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.
- [Luc *et al.*, 2020a] Pauline Luc, Aidan Clark, Sander Dieleman, Diego de Las Casas, Yotam Doron, Albin Cassirer, and Karen Simonyan. Transformation-based adversarial video prediction on large-scale data. *arXiv preprint arXiv:2003.04035*, 2020.
- [Luc *et al.*, 2020b] Pauline Luc, Aidan Clark, Sander Dieleman, Diego de Las Casas, Yotam Doron, Albin Cassirer, and Karen Simonyan. Transformation-based adversarial video prediction on large-scale data. *CoRR*, abs/2003.04035, 2020.
- [Lugmayr *et al.*, 2022] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

- [Moing *et al.*, 2021] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. CCVS: context-aware controllable video synthesis. In Marc’ Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, pages 14042–14055, 2021.
- [Muñoz *et al.*, 2021] Andrés Muñoz, Mohammadreza Zolfaghari, Max Argus, and Thomas Brox. Temporal shift GAN for large scale video generation. In *IEEE Winter Conference on Applications of Computer Vision*, pages 3178–3187, 2021.
- [Nichol and Dhariwal, 2021] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.
- [Nichol *et al.*, 2021] 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.
- [Niklaus *et al.*, 2017] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 670–679, 2017.
- [Rakhimov *et al.*, 2021] Ruslan Rakhimov, Denis Volkonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 101–112, 2021.
- [Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [Rombach *et al.*, 2022] 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*, pages 10684–10695, 2022.
- [Saharia *et al.*, 2022] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasempour, Burcu Karagol Ayan, Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.
- [Salimans and Ho, 2022] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.
- [Singer *et al.*, 2022] 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.
- [Song *et al.*, 2020a] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [Song *et al.*, 2020b] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.
- [Tran *et al.*, 2015] 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*, pages 4489–4497, 2015.
- [Tulyakov *et al.*, 2018] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1526–1535, 2018.
- [Unterthiner *et al.*, 2018] 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.
- [Voleti *et al.*, 2022] Vikram Voleti, Alexia Jolicœur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. *arXiv preprint arXiv:2205.09853*, 2022.
- [Vondrick *et al.*, 2016] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, pages 613–621, 2016.
- [Weissenborn *et al.*, 2019] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019.
- [Wu *et al.*, 2021] 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*, pages 2318–2328, 2021.
- [Xu *et al.*, 2020] Qiangeng Xu, Hanwang Zhang, Weiyue Wang, Peter Belhumeur, and Ulrich Neumann. Stochastic dynamics for video infilling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2714–2723, 2020.
- [Yan *et al.*, 2021] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using VQ-VAE and transformers. *CoRR*, abs/2104.10157, 2021.
- [Yang *et al.*, 2022] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022.
- [Yu *et al.*, 2022] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022.