Latent Video Diffusion Models for High-Fidelity Long Video Generation

Yingqing He[1]   Tianyu Yang[2]   Yong Zhang[2]   Ying Shan[2]   Qifeng Chen[1]
[1]The Hong Kong University of Science and Technology   [2]Tencent AI Lab

Abstract

AI-generated content has attracted lots of attention recently, but photo-realistic video synthesis is still challenging. Although many attempts using GANs and autoregressive models have been made in this area, the visual quality and length of generated videos are far from satisfactory. Diffusion models have shown remarkable results recently but require significant computational resources. To address this, we introduce lightweight video diffusion models by leveraging a low-dimensional 3D latent space, significantly outperforming previous pixel-space video diffusion models under a limited computational budget. In addition, we propose hierarchical diffusion in the latent space such that longer videos with more than one thousand frames can be produced. To further overcome the performance degradation issue for long video generation, we propose conditional latent perturbation and unconditional guidance that effectively mitigate the accumulated errors during the extension of video length. Extensive experiments on small domain datasets of different categories suggest that our framework generates more realistic and longer videos than previous strong baselines. We additionally provide an extension to large-scale text-to-video generation to demonstrate the superiority of our work. Our code and models will be made publicly available.

1. Introduction

A video can provide more informative, attractive, and immersive visual content that presents the physical 3D world. A powerful video generation tool can benefit novel content creation, gaming, and movie production. Thus, rendering photorealistic videos is a long-standing and exciting goal of the computer vision and graphics research community. However, the high-dimensional video samples and the statistical complexity of real-world video distributions make video synthesis quite challenging and computationally expensive.

Existing works capitalize on different types of generative models including GANs [40, 26, 27, 36, 35, 30, 2], VAEs [7, 44], autoregressive models [6, 42, 21], and nor-

malizing flows [17]. Particularly, GANs have achieved great success in image generation [1, 15, 16, 14], thus extending it to video generation with a dedicated temporal design achieves outstanding results [35, 30]. However, GANs suffer from mode collapse and training instability problems, which makes GAN-based approaches hard to scale up to handle complex and diverse video distributions. Most recently, TATS [6] proposes an autoregressive approach that leverages the VQGAN [5] and transformers to synthesize long videos. However, the generation fidelity and resolution (128×128) still have much room for improvement.

To overcome these limitations, we leverage diffusion models (DMs) [9], another class of generative models that achieve impressive performance in various image synthesis tasks [4, 22, 18, 25, 23]. However, directly extending DMs to video synthesis requires substantial computational resources [12, 39]. Besides, most of the text-to-video generation models are not available to the public [8, 29, 48, 12], which hinders the research progress of this field. To tackle these problems, we devise LVDM, an efficient video diffusion model in the latent space of videos and we achieve state-of-the-art results via the simple base LVDM model. In addition, to further generate long-range videos, we introduce a hierarchical LVDM framework that can extend videos far beyond the training length. However, generating long videos tends to suffer the performance degradation problem. To mitigate this issue, we propose conditional latent perturbation and unconditional guidance, which effectively slow the performance degradation over time. Ultimately, our framework surpasses many previous works in short and long video generation and establishes new state-of-the-art performance on three datasets. We also provide additional results for text-to-video generation as an extension. Notably, our codes and pre-trained models will be publicly available as an efficient diffusion baseline model for video synthesis and downstream video editing tasks.
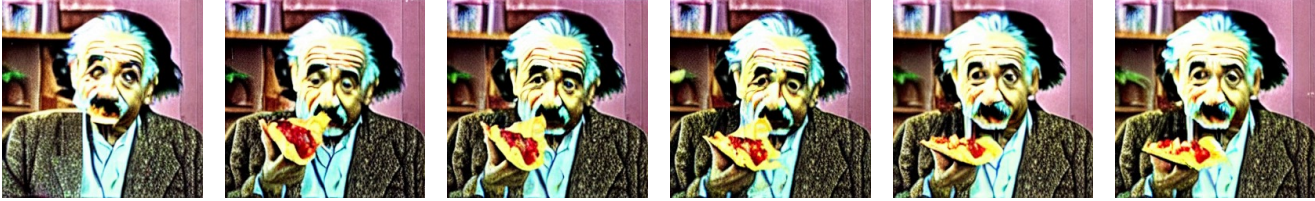
In sum, our work makes the following contributions:

- We introduce LVDM, an efficient diffusion-based baseline approach for video generation by firstly compressing videos into tight latents.

- We propose a hierarchical framework that operates in
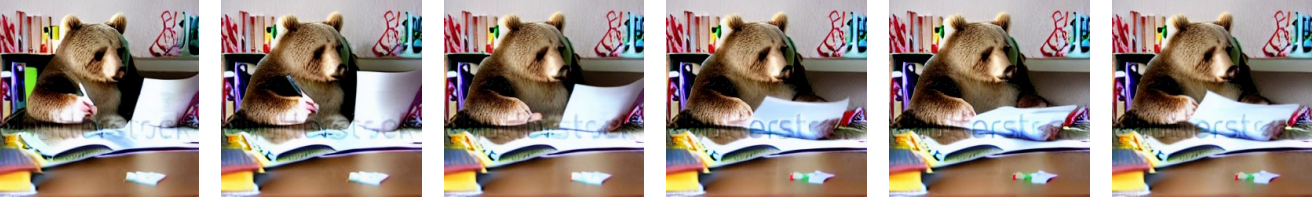
"astronaut riding a horse"
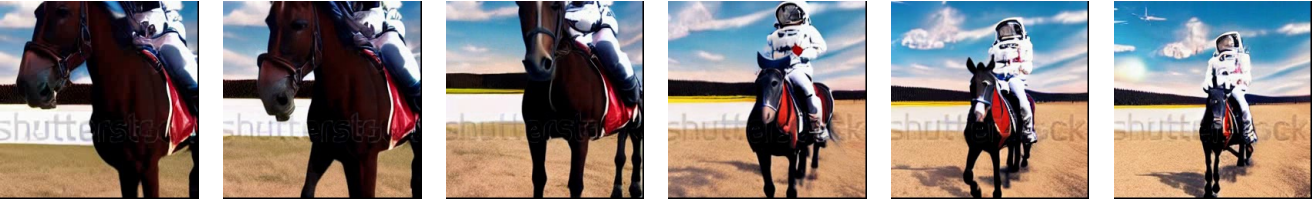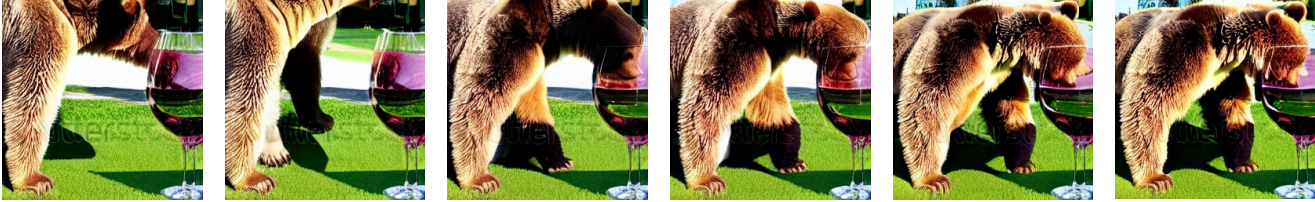

" Albert Einstein eating pizza"

"A happy elephant wearing a big birthday hat walking under the sea, 4k"

"A teddy bear is writing paper"

"A teddy bear is drinking a big wine"

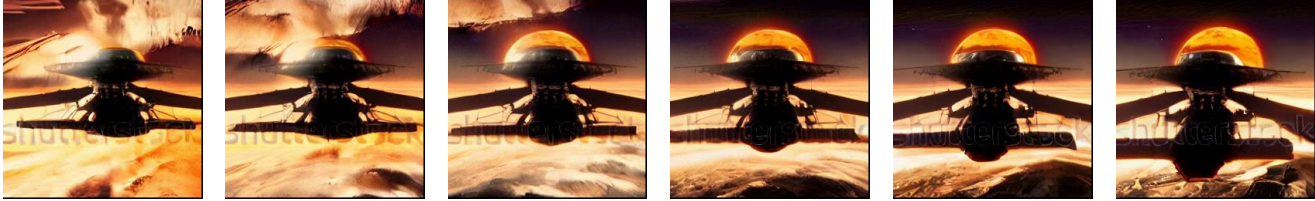"A giant spaceship is landing on mars in the sunset. High Definition."

Figure 1: Results of extending our LVDM to text-to-video generation.

the video latent space, enabling our models to generate longer videos beyond the training length further.

- We propose *conditional latent perturbation* and *unconditional guidance* for mitigating the performance degradation issue during long video generation.

- Our model achieves state-of-the-art results on three benchmarks in both short and long video generation settings. We also provide appealing results for open-domain text-to-video generation, demonstrating the effectiveness and generalization of our models.
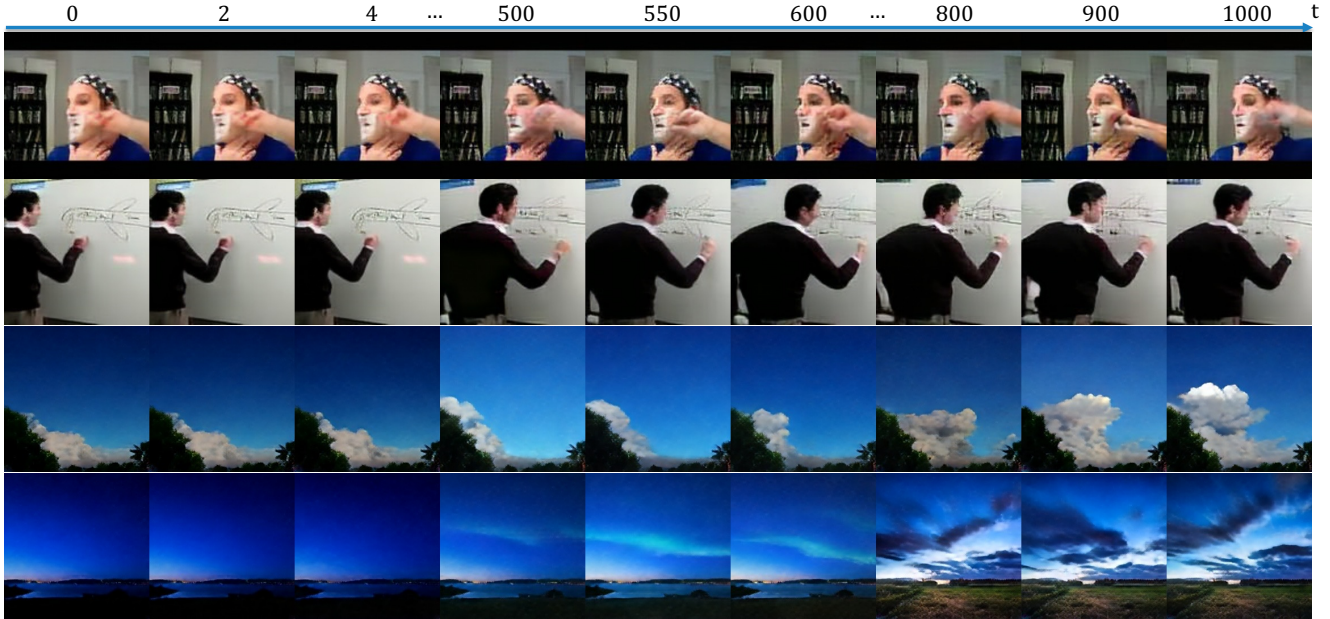
Figure 2: Our unconditional long video generation results on UCF-101 [34] and Sky Time-lapse [43]. t indicates the frame index.

2. Related Work

2.1. Video Synthesis

Video synthesis aims to model the distribution of real-world videos, and then one can randomly draw realistic and novel video samples from the learned distribution. Prior works mainly exploit deep generative models, including GANs [46, 30, 35, 36, 27, 26, 40], autoregressive models [44, 6], VAEs [41, 7], and normalizing flows [17]. Among them, the most dominant ones are GAN-based approaches due to the great success of GANs in image modeling. MoCoGAN [36] and MoCoGAN-HD [35] learn to decompose latent codes into two subspaces, i.e., content and motion. MoCoGAN-HD [35] leverages the powerful pretrained StyleGAN2 as the content generator, demonstrating higher-resolution video generation results. StyleGAN-V [30] and DiGAN [46] introduce implicit neural representation to GANs for modeling the continuity of temporal dynamics. They built long-video GAN on top of StyleGAN3 and apply hierarchical generator architecture for long-range modeling, thus producing videos with new content arising in time. Despite these achievements made by GANs, those methods tend to suffer from mode collapse and training instability. Autoregressive methods have also been exploited for video generation. VideoGPT [44] uses VQVAE [20] and transformer to autoregressively generate tokens in a discrete latent space. TATS [6] changes the VQVAE [20] to a more powerful VQGAN [5] and combines a frame interpolation transformer to render long videos in a hierarchical manner. Different from the aforementioned methods, we study diffusion models for video generation in this work.

2.2. Diffusion Models

Diffusion models are a class of likelihood-based generative models that have shown remarkable progress in image synthesis tasks. Due to their desirable properties like stable training and easy scalability, diffusion models have surpassed GANs on the image generation task and achieved both higher sample quality and better distribution coverage [4]. Its scalability further facilitates the building of large-scale text-to-image generation models [25, 18, 23, 22], which show marvelous image samples recently. The groundbreaking work of diffusion models is DDPM [9], which leverages the connection between Langevin dynamics [31] and denoising score matching [33] to build a weighted variational bound for optimization. However, the sampling process needs to follow the Markov chain step by step to produce one sample, which is extremely slow (e.g., usually 1000 or 4000 steps [9, 4]). DDIM [32] accelerates the sampling process via an iterative non-Markovian way while keeping the same training process unchanged. [19] further improves the log-likelihoods while maintaining high sample quality, and ADM [4] eventually outperforms GAN-based methods via an elaborately designed architecture and classifier guidance. Contemporarily, a cascaded diffusion model [10] is proposed with a sequence of lower resolution conditional diffusion models as an alternative approach to improve the sample fidelity and shows that conditional noisy augmentation is pivotal to the stacked architecture. Despite their findings and achievement in image synthesis, diffusion models on video generation have not been well-studied. Most recently,
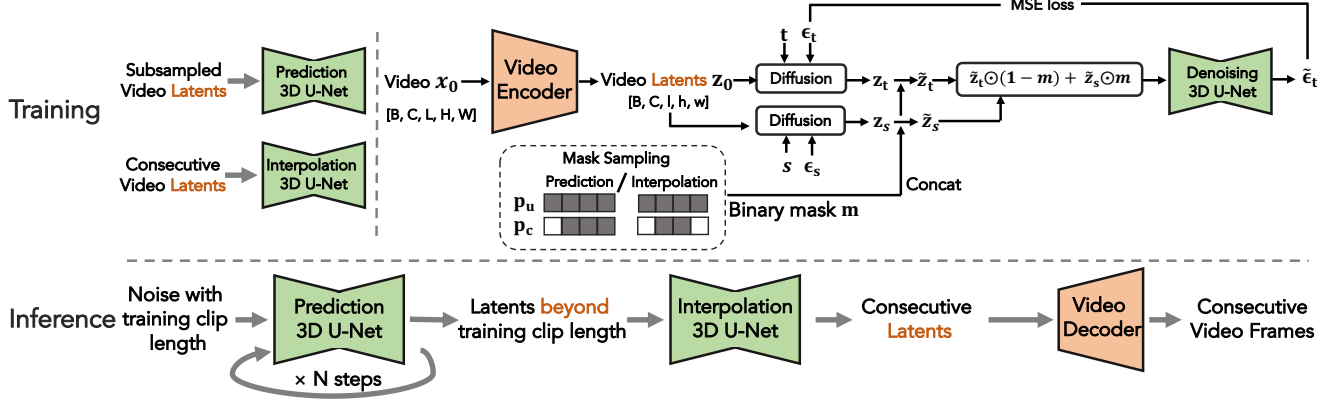
Figure 3: **Hierarchical LVDM Framework.** We present a hierarchical latent video diffusion model (LVDM) for generating longer videos beyond the temporal training length. t and s are randomly sampled diffusion timesteps for generated latents and conditional latents, respectively. p_c and p_u are probabilities of the conditional and unconditional input, respectively.

VDM [12] extends diffusion models to the video domain, which initiates the exploration of diffusion models on video generation. Specifically, they modify the 2D UNet to a spatial-temporal factorized 3D network and further present image-video joint training and gradient-based video extension techniques. MCVD [39] parameterizes unconditional video generation and conditional frame prediction and interpolation models as a unified one by randomly dropping conditions (previous or future frames) during training, which shares the similar underneath idea with classifier-free guidance which joint trains a class-conditional and unconditional model. Make-A-Video [29], and Imagen Video [8] leverage a series of big diffusion models for large-scale video synthesis conditioned on a given text prompt. However, previous video-based video generation approaches all perform diffusion and denoising processes in pixel space, which requires substantial computational resources. In this paper, we extend the latent image diffusion model [23] to video by devising a 3D auto-encoder for video compression. Founded on this baseline, we further show how long videos can be sampled via a hierarchical architecture and natural extensions of conditional noise augmentation.

Concurrent works. With the rapid development of diffusion models, two contemporary works share similar ideas with us and propose latent-based diffusion models. MagicVideo [48] proposes an efficient video generation framework via latent diffusion models. Unlike MagicVideo, which compresses videos frame-by-frame, we compress the redundant information along the temporal axis to obtain more compact latent. PVDM [45] proposes to parameterize videos as image-like latent by selecting three 2D latents along three axes of the 3D video latent. Then they train a 2D diffusion network to model these 2D latents for video. Differently, we exploit 3D diffusion networks to model the cubic video latents directly. Another important difference between our work with Magic video and PVDM is that we

make a further step towards long video generation and provide a hierarchical framework, conditional noise perturbation and unconditional guidance to boost the long video generation performance. on alleviating the performance degradation problem.

3. Method

We firstly compress video samples to a lower-dimensional latent space by a video autoencoder. Then we design a unified video diffusion model, which can perform both unconditional generation and conditional video generation in one network, in the latent space. This enables our model to self-extend the generated video to an arbitrary length autoregressively. However, autoregressive models only tend to suffer the problem of performance degradation over time. To further improve the coherence of generated long video and alleviate the quality degradation problem, we propose hierarchical latent video diffusion models to first generate video patents sparsely and then interpolate intermediate latents. We also propose conditional latent perturbation and unconditional guidance for tackling the performance degradation problem in long video generation.

3.1. Video Autoencoder

We compress videos using a lightweight 3D autoencoder, including an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. Both of them consist of several layers of 3D convolutions. Formally, given a video sample $\mathbf{x}_0 \sim p_{data}(\mathbf{x}_0)$ where $\mathbf{x}_0 \in \mathbb{R}^{H \times W \times L \times 3}$, the encoder $\mathcal{E}$ encodes it to its latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$ where $\mathbf{z}_0 \in \mathbb{R}^{h \times w \times l \times c}$, $h = H/f_s$, $w = W/f_s$, and $l = L/f_t$. f_s and f_t are spatial and temporal downsampling factors. The decoder $\mathcal{D}$ decodes $\mathbf{z}_0$ to the reconstructed sample $\tilde{\mathbf{x}}_0$, i.e. $\tilde{\mathbf{x}}_0 = \mathcal{D}(\mathbf{z}_0)$. To ensure that the autoencoder is temporally shift-equivariant, we follow [6] to use repeat padding in all three-dimensional convolutions. The training objective includes a reconstruction loss $\mathcal{L}_{rec}$ and an

adversarial loss $\mathcal{L}_{adv}$. The reconstruction loss $\mathcal{L}_{rec}$ is comprised of a pixel-level mean-squared error (MSE) loss and a perceptual-level LPIPS [47] loss. The adversarial loss [5] is used to eliminate the blur in reconstruction usually caused by the pixel-level reconstruction loss and further improve the realism of the reconstruction. In summary, the overall training objective of $\mathcal{E}$ and $\mathcal{D}$ is

$$\mathcal{L}_{AE} = \min_{\mathcal{E},\mathcal{D}} \max_{\psi} \big(\mathcal{L}_{rec}(\mathbf{x}_0, \mathcal{D}(\mathcal{E}(\mathbf{x}_0))) \\ +\mathcal{L}_{adv}(\psi(\mathcal{D}(\mathcal{E}(\mathbf{x}_0))))\big). \quad (1)$$

where ψ is the discriminator used in adversarial training.

3.2. Base LVDM for Short Video Generation

Revisiting Diffusion Models. We propose to perform diffusion and denoising on the video latent space. Given a compressed latent code $\mathbf{z}_0 \sim p_{data}(\mathbf{z}_0)$, we train diffusion models to generate latent samples starting from a pure Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{z}_T; \mathbf{0}, \mathbf{I})$ in T timesteps, producing a set of noisy latent variables, i.e., $\mathbf{z}_1, ..., \mathbf{z}_T$. The forward diffusion process is gradually adding noise to $\mathbf{z}_0$ according to a predefined variance schedule $\beta_1, \ldots, \beta_T$:

$$q(\mathbf{z}_{1:T}|\mathbf{z}_0) \coloneqq \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{z}_{t-1}), \quad (2)$$

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) \coloneqq \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}). \quad (3)$$

Eventually, the data point $\mathbf{z}_T$ becomes indistinguishable from pure Gaussian noise. To recover $\mathbf{z}_0$ from $\mathbf{z}_T$, diffusion models learn a backward process via

$$p_\theta(\mathbf{z}_{0:T}) \coloneqq p(\mathbf{z}_T) \prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t), \quad (4)$$

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) \coloneqq \mathcal{N}(\mathbf{z}_{t_1}; \mu_\theta(\mathbf{z}_t, t), \Sigma_\theta(\mathbf{z}_t, t)), \quad (5)$$

where θ is a parameterized neural network, typically a U-Net [24] commonly used in image synthesis, to predict $\mu_\theta(\mathbf{z}_t, t)$, and $\Sigma_\theta(\mathbf{z}_t, t)$. In practice, we parameterize $\mu_\theta(\mathbf{z}_t, t)$ by

$$\mu_\theta(\mathbf{z}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{z}_t, t)\right), \quad (6)$$

where $\epsilon_\theta(\mathbf{z}_t, t)$ is eventually estimated. We simply set $\Sigma_\theta(\mathbf{z}_t, t) = \beta_t^2\mathbf{I}$ as in [9].

The training objective is a simplified version of variational bound:

$$\mathcal{L}_{\text{simple}}(\theta) \coloneqq \|\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t) - \boldsymbol{\epsilon}\|_2^2, \quad (7)$$

where $\boldsymbol{\epsilon}$ is drawn from a diagonal Gaussian distribution.
Video Generation Backbone. To model video samples in the 3D latent space, we follow [12] that exploits a spatial-temporal factorized 3D U-Net architecture to estimate the $\boldsymbol{\epsilon}$. Specifically, we use space-only 3D convolution with the shape of $1 \times 3 \times 3$ and add temporal attention in partial layers. We investigate two kinds of attention: joint spatial-temporal self-attention and factorized spatial-temporal self-attention. We observe that applying the joint spatial-temporal attention does not exhibit significant benefit compared with the factorized one while increasing the model complexity and introducing spot-like artifacts in random locations sometimes. Thus we use factorized spatial-temporal attention as the default setting in our experiments. We use adaptive group normalization to inject the timestep embedding into normalization modules to control the channel-wise scale and bias parameters, which have also been demonstrated to be beneficial for improving sample fidelity in [4].

3.3. Hierarchical LVDM for Long Video Generation

The aforementioned framework can only generate short videos, whose lengths are determined by the input frame number during training. We therefore propose a conditional latent diffusion model, which can produce future latent codes conditioned on the previous ones in an autoregressive manner, to facilitate long video generation. We further present several techniques to alleviate the error accumulation problem in autoregressive generation.
Autoregressive Latent Prediction. Considering a short clip latent $\mathbf{z}_t = \{\mathbf{z}_t^i\}_{i=i}^l$ where $\mathbf{z}_t^i \in \mathbb{R}^{h \times w \times c}$ and l is the number of latent codes within the clip, we can learn to predict future latent codes conditioned on the former ones. For each video frame in a clip latent, we add an additional binary mask along the channel dimension to indicate whether it is a conditional frame or a frame to predict and replace the $\mathbf{z}_t^i$ with $\mathbf{z}_0^i$ according to the mask, yielding,

$$\tilde{\mathbf{z}}_t = \{\tilde{\mathbf{z}}_t^i = [\mathbf{z}_t^i, \mathbf{m}^i]\}_{i=1}^l\}_{i=1}^l, \tilde{\mathbf{z}}_t^i \in \mathbb{R}^{h \times w \times (c+1)}$$
$$\tilde{\mathbf{z}}_0 = \{\tilde{\mathbf{z}}_0^i = [\mathbf{z}_0^i, \mathbf{m}^i]\}_{i=1}^l\}_{i=1}^l, \tilde{\mathbf{z}}_0^i \in \mathbb{R}^{h \times w \times (c+1)} \quad (8)$$
$$\tilde{\mathbf{z}}_t \leftarrow \tilde{\mathbf{z}}_t \odot (1 - \mathbf{m}) + \tilde{\mathbf{z}}_0 \odot \mathbf{m}$$

where $\mathbf{m} = \{\mathbf{m}^i\}_{i=1}^l, \mathbf{m}^i \in \mathbb{R}^{h \times w \times 1}$ is the binary mask. By randomly setting different binary masks to ones or zeros, we can train our diffusion model to perform both unconditional video generation and conditional video prediction jointly. Concretely, we set all masks in the binary clip $\mathbf{m}$ to zeros for unconditional diffusion model training. During inference stage, we set the first k binary mask $\{\mathbf{m}^i\}_{i=1}^k$ to ones and the remaining $\{\mathbf{m}^i\}_{i=k+1}^l$ to zeros.
Hierarchical Latent Generation. Generating videos in an autoregressive way has the risk of quality degradation caused by accumulated error over time. We thus utilize a common strategy, hierarchical generation [6, 13, 3], to alleviate this problem. Specifically, we first train an autoregressive video generation model on *sparse frames* to form the

basic storyline of the video and then train another interpolation model to infill the missing frames. The training of the interpolation model is similar to the autoregressive model, except that we set the binary masks of the middle frames between every two sparse frames to zeros.

Conditional Latent Perturbation. Although the above-mentioned hierarchical generation manner can reduce the number of autoregressive steps to overcome the degradation issue, more prediction steps are indispensable to produce long-enough video samples. Thus, we propose conditional perturbation to mitigate the conditional error induced by the previous generation step. Specifically, rather than directly conditioning on $\mathbf{z}_0$, we use the noisy latent code $\mathbf{z}_s$ at an arbitrary time s, which could be computed by (3), as the condition during training, i.e., $\{\mathbf{z}_t^i\}_{i=1}^k \leftarrow \{\mathbf{z}_s^i\}_{i=1}^k$. This means we also perform a diffusion process on the conditional frames. To keep the conditional information preserved, a maximum threshold s_{max} is used to clamp the timesteps in a minor noise level. During sampling, a fixed noise level is used to consistently add noise during autoregressive prediction. Conditional latent perturbation is inspired by conditional noise augmentation, which has been proposed in cascaded diffusion models [10] to improve the performance of super-resolution diffusion models. We extend it to video prediction, and we are the first to demonstrate its effectiveness in producing long video samples.

Unconditional Guidance. Another complementary technique to alleviate quality degradation of autoregressive video generation is to leverage the unconditional score to guide the conditional generation process. Since the accumulated error during autoregressive generation does not affect the unconditional score, introducing this score into long video generation could improve the diversity and fidelity of sampled video. Thanks to the joint training techniques presented above, we can use one network to estimate both unconditional scores $\boldsymbol{\epsilon}_u$ and conditional scores $\boldsymbol{\epsilon}_c$. By zeroing all binary maps $\{\mathbf{m}^i\}_{i=1}^l$ in $\tilde{\mathbf{z}}$, we obtain $\boldsymbol{\epsilon}_u$. By setting the first k binary maps $\{\mathbf{m}^i\}_{i=1}^k$ to one and the remaining ones $\{\mathbf{m}^i\}_{i=k+1}^l$ to zero in $\tilde{\mathbf{z}}$, we get $\boldsymbol{\epsilon}_c$. Note that the conditional score may be out of the model learned distribution due to the error accumulation when autoregressively predicted. Thus we propose to leverage the unconditional scores to guide the prediction sampling process via

$$\tilde{\epsilon}_\theta = (1 + w)\boldsymbol{\epsilon}_c - w\boldsymbol{\epsilon}_u, \tag{9}$$

where w is the guidance strength. This formula is initially presented in [11] and referred to as *classifier-free guidance* to avoid training a separate classifier for class-conditional diffusion models. We extend this idea to guide the frame-conditional diffusion models to generate longer videos.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our method on UCF-101 [34], Sky Time-lapse [43], and Taichi [28]. We train all our models with the resolution of 256^2 on these datasets for unconditional video generation. The short clips used for training are selected with the frame stride of 1 at a random location of one video. For taichi, we select clips with the frame stride of 4 (i.e., skip three frames after selecting one) following prior work [6, 46] to make the human motion more dynamic. Due to the limited number of videos in UCF-101 and TaiChi, we adopt the full dataset for training. For the Sky Time-lapse dataset, we only train the model on its training split. All models are trained under the unconditional setting with no guidance information provided, such as class label.

Evaluation Metrics. For quantitative evaluation, we report the commonly-used FVD [37] and KVD [38] for both short and long video generation. Specifically, we calculate FVD and KVD between 2048 real and fake videos with 16 frames, which we refer to as FVD_{16} and KVD_{16}, for short video evaluation. All results for short video generation are calculated among ten runs and report their mean and standard deviation. For long video evaluation, we estimate FVD and KVD among 512 samples in every non-overlapped 16-frame clip and report an FVD curve across 1024 frames calculated in 1 run, referred to as FVD_{1024}.

Baselines. We compared our approach with seven competing baselines, including GAN-based methods: TGAN-v2 [27], DiGAN [46], MoCoGAN-HD [35], and long-video GAN [2], autoregressive models TATS [6], and most recent diffusion-based models including Video Diffusion Models (VDM) [12] and MCVD [39]. Results are collected from their original papers except for VDM. Please note that VDM has mentioned in its main paper that it will not publish the source code. Thus we implement it by ourselves. Detailed implementations are documented in the supplement.

Implementation Details. We use the spatial and temporal downsampling factors of 8 and 4, respectively. We first train a 3D autoencoder, then fix its weights, and then start to train an unconditional LVDM on short video clips. After that, the LVDM-prediction and LVDM-interpolation models are resumed from the unconditional one. More details are illustrated in supplementary materials.

4.2. Efficiency Comparison

To demonstrate the training and sampling efficiency of our method, we compare our approach with two pixel-space video diffusion models, including VDM [12] and MCVD [39] in Tab. 2 on the UCF-101 dataset. We implemented VDM with a base unconditional video diffusion model to synthesize videos of low resolution (16 frames

Figure 4: Qualitative comparison with state-of-the-art methods on unconditional short video generation (16 frames) on three datasets. All frames shown are selected at $t = 10$ from four *randomly* generated samples.

Method	$FVD_{16} \downarrow$	$KVD_{16} \downarrow$
Resolution 128^2		
MoCoGAN-HD [35]	183.6 ± 5.2	13.9 ± 0.7
DIGAN [46]	114.6 ± 4.9	6.8 ± 0.5
TATS [6]	132.6 ± 2.6	5.7 ± 0.3
Long-video GAN [6]	107.5	-
Resolution 256^2		
Long-video GAN [6]	116.5	-
LVDM (Ours)	$\mathbf{95.2 \pm 2.3}$	$\mathbf{3.9 \pm 0.1}$

(a) Sky Time-lapse

Method	$FVD_{16} \downarrow$	$KVD_{16} \downarrow$
Resolution 64^2		
MCVD [39]	1143	-
Resolution 128^2		
TGAN-v2 [27]	1209 ± 28	-
DIGAN* [46]	577 ± 21	-
TATS [6]	420 ± 18	-
Resolution 256^2		
MoCoGAN-HD* [35]	700 ± 24	-
LVDM (Ours*)	$\mathbf{372 \pm 11}$	27 ± 1

(b) UCF-101

Method	$FVD_{16} \downarrow$	$KVD_{16} \downarrow$
Resolution 128^2		
MoCoGAN-HD [35]	144.7 ± 6.0	25.4 ± 1.9
DIGAN [46]	128.1 ± 4.9	20.6 ± 1.1
TATS [6]	$\mathbf{94.6 \pm 2.7}$	$\mathbf{9.8 \pm 1.0}$
Resolution 256^2		
DIGAN [46]	156.7 ± 6.2	-
LVDM (Ours)	$\mathbf{99.0 \pm 2.6}$	$\mathbf{15.3 \pm 0.9}$

(c) Taichi

Table 1: Quantitative comparisons of short video generation on three datasets.

Method	#Dims	#Params	Step Time	$FVD_{16} \downarrow$	$KVD_{16} \downarrow$
VDM [12]	197k	445M (397+48)	4.9s (3.3+1.6)	1396	116
MCVD [39]	786k	441M	3s	2460	148
LVDM (Ours)	16k	437M (418+19)	0.8s	**552**	**42**

Table 2: Efficiency and performance comparisons between our latent-space approach with two pixel-space approaches, VDM and MCVD, on unconditional short video generation on UCF-101. All models are trained with approximately the same parameters, the same training time (4.5 days), and the same devices (8 V100s). + indicates two models needed.

with the resolution of 64^2) and a video super-resolution diffusion model to upscale the spatial resolution to 256^2. We train MCVD following its official setting, except that we scale it to resolution 256^2. Our method achieves better FVD than them when trained with a similar time and number of model parameters.

4.3. Short Video Generation

Quantitative Results. In Tab. 1, we provide quantitative comparisons with previous methods on Sky-Timelapse, UCF-101, and Taichi, respectively. Our method outperforms previous state-of-the-art methods by a large margin. Specifically, on the Sky-Timelapse dataset, we reduce FVD

from 116.5 to 95.18 under the resolution of 256^2. In addition, our high-resolution performance also surpasses those of the state-of-the-art methods in FVD under the resolution of 128^2. On UCF-101 and taichi datasets, we achieve new state-of-the-art results under the resolution of 256^2, while our FVD is comparable to the best result under the resolution of 128^2. LVDM outperforms diffusion-based method MCVD on UCF-101 regarding both FVD and resolution.

Qualitative Results. In Fig. 4, we showcase visual comparisons with DIGAN [46] and TATS [6]. We observe that samples produced by DIGAN exhibit coordinates-like artifacts in many samples, TATS tends to generate samples with flat contents and lacks diversity, while our LVDM can synthesize video samples with high fidelity and diversity. More comparisons are documented in supplementary materials.

4.4. Long Video Generation

Comparison with State-of-the-art Approach. We compare our method LVDM with TATS [6] for long video generation with 1024 frames on the UCF-101 and Sky Time-lapse datasets. Both LVDM and TATS experiment with a pure autoregressive prediction model (autoregressive) and a combination of a prediction model with an interpolation model (hierarchical). Fig. 5 and Fig. 6 present the qualitative and quantitative comparison results, respectively. Note

Figure 5: Qualitative results of generated long videos compared with state-of-the-art approach TATS[6] on UCF-101. Each frame is selected with a frame step 16. Both approaches are compared in the autoregressive setting.
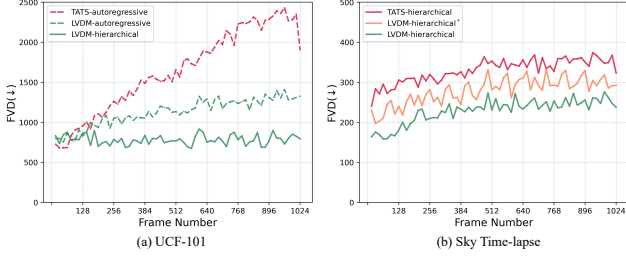

Figure 6: Quantitative comparison with TATS on long video generation (1024 frames) on two datasets. LVDM outperforms TATS in both autoregressive and hierarchical manners. * indicates DDIM sampling with 100 timesteps.

that TATS does not provide its hierarchical checkpoints on UCF-101. On the Sky Time-lapse, we compare two methods under a hierarchical setting. Our method achieves both lower FVD scores and slower quality degradation over time compared with TATS. Specifically, on challenging UCF-101, our LVDM-autoregressive has much slower quality degradation over time than TATS-autoregressive. Our LVDM-hierarchical further significantly alleviates the quality degradation problem. On Sky Timelapse, both approaches achieve minor quality degradation over time, while our methods have better generation performance.

Conditional Latent Perturbation. Fig. 7 (a) shows the ablation results of conditional latent perturbation. We can observe that it can slow down the performance degradation trend, especially when the frame number is larger than 512, demonstrating its effectiveness on long video generation.

Unconditional Guidance. We also explore *unconditional guidance* attempting to alleviate quality degradation of autoregressive generation by leveraging the unconditional score to guide the autoregressive conditional generation with a guidance strength. Fig. 7 (b) demonstrates that this technique eases the quality degradation of long video generation effectively.

5. Extension for Text-to-Video Generation

Besides unconditional video generation, we extend our approach to more controllable text-to-video generation. We scale up our model to a billion-level of parameters and train
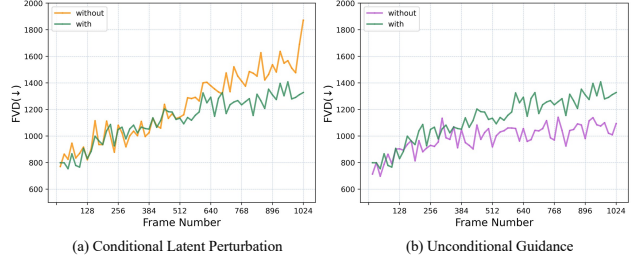

Figure 7: Ablation experiments of conditional latent perturbation and unconditional guidance.

our model on a 2 million subset of the WebVid dataset. To efficiently learn video generation models, one natural idea is to leverage the pre-trained text-to-image models to reuse the spatial content generation capacity and continue to learn motion dynamics on the video dataset. Thus, we initialize the spatial parameters of our model, including the spatial convolutions and spatial attention, using the pre-trained stable diffusion weights. And we initialize the temporal modules, including the multi-layer temporal self-attention, as identity mapping to preserve the spatial generation performance. The text-to-video results are shown in Fig. 1. More video results are documented in supplementary materials.

6. Conclusion

In this work, we devise an efficient DM-based framework for video generation, which significantly reduces the data dimension and speeds up the training and sampling. Our hierarchical framework can effectively generate long videos with more than one thousand of frames. With its better modeling capacity, we achieve new state-of-the-art results on various datasets under short and long video generation settings. We further demonstrate the effectiveness of unconditional guidance and conditional latent perturbation in reducing the accumulated error induced during autoregressively extending video lengths. We also provide an extension of our model for open-domain text-to-video generation. We hope our method could serve as a strong baseline for future video generation works. Future explorations could be made regarding better architecture design choices and further speeding up the training and sampling of video diffusion models.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 1

[2] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A Efros, and Tero Karras. Generating long videos of dynamic scenes. *arXiv preprint arXiv:2206.03429*, 2022. 1, 6

[3] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Hierarchical video generation for complex data. *arXiv preprint arXiv:2106.02719*, 2021. 5

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 3, 5

[5] 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*, pages 12873–12883, 2021. 1, 3, 5

[6] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. *arXiv preprint arXiv:2204.03638*, 2022. 1, 3, 4, 5, 6, 7, 8

[7] Jiawei He, Andreas Lehrmann, Joseph Marino, Greg Mori, and Leonid Sigal. Probabilistic video generation using holistic attribute control. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018. 1, 3

[8] 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. 1, 4

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 3, 5

[10] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022. 3, 6

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[12] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 1, 4, 5, 6, 7

[13] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 5

[14] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *NeurIPS*, 2021. 1

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 1

[17] Manoj Kumar, Mohammad Babaeizadeh, Dumitru Erhan, Chelsea Finn, Sergey Levine, Laurent Dinh, and Durk Kingma. Videoflow: A conditional flow-based model for stochastic video generation. *arXiv preprint arXiv:1903.01434*, 2019. 1, 3

[18] 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

[19] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[20] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 3

[21] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. *arXiv preprint arXiv:2006.10704*, 2020. 1

[22] 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. 1, 3

[23] 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. 1, 3, 4

[24] 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*, pages 234–241. Springer, 2015. 5

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 3

[26] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, pages 2830–2839, 2017. 1, 3

[27] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *International Journal of Computer Vision*, 128:2586–2606, 2020. 1, 3, 6, 7

[28] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 2019. 6

[29] 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. 1, 4

[30] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3626–3636, 2022. 1, 3

[31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[33] 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

[34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3, 6

[35] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *International Conference on Learning Representations*, 2021. 1, 3, 6, 7

[36] 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*, pages 1526–1535, 2018. 1, 3

[37] 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. 6

[38] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *ICLR*, 2019. 6

[39] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Masked conditional video diffusion for prediction, generation, and interpolation. *arXiv preprint arXiv:2205.09853*, 2022. 1, 4, 6, 7

[40] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016. 1, 3

[41] Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with vqvae. *arXiv preprint arXiv:2103.01950*, 2021. 3

[42] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019. 1

[43] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3, 6

[44] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 1, 3

[45] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. *arXiv preprint arXiv:2302.07685*, 2023. 4

[46] 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. 3, 6, 7

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[48] 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. 1, 4