MoVideo: Motion-Aware Video Generation with Diffusion Model

Jingyun Liang[1], Yuchen Fan[2], Kai Zhang[4⋆], Radu Timofte[3], Luc Van Gool[1], and Rakesh Ranjan[2]

[1] Computer Vision Lab, ETH Zurich, Switzerland
[2] Meta Inc.
[3] Computer Vision Lab, CAIDAS & IFI, University of Würzburg, Germany
[4] Nanjing University, Suzhou, China
{jinliang,vangool}@vision.ee.ethz.ch radu.timofte@uni-wuerzburg.de
{ycfan, rakeshr}@meta.com kaizhang@nju.edu.cn
https://jingyunliang.github.io/MoVideo

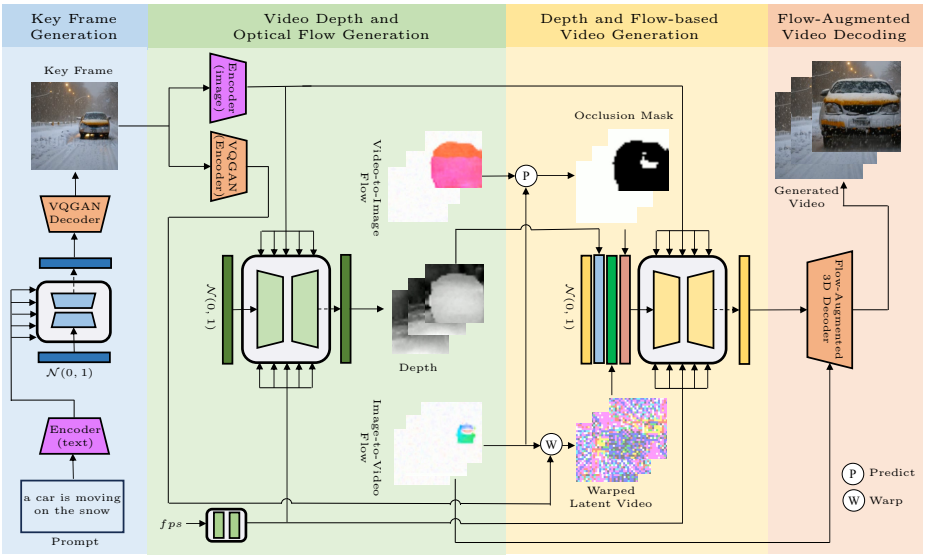

Fig. 1: The schematic illustration of the proposed motion-aware video generation (MoVideo) framework. Given a text prompt, we first generate the key frame by a public available latent diffusion model. Then, we generate video depth and optical flows conditional on the image embedding (extracted by an open-sourced pretrained image-text bi-encoder model) and frames per second. Next, we add extra conditions, including depth, flow-based warped latent video and calculated occlusion mask, to generate the video in the latent space. Last, the video is decoded with flow-based alignment and feature refinement modules.

Abstract. While recent years have witnessed great progress on using diffusion models for video generation, most of them are simple exten-

⋆ Corresponding author.

sions of image generation frameworks, which fail to explicitly consider one of the key differences between videos and images, *i.e.*, motion. In this paper, we propose a novel motion-aware video generation (MoVideo) framework that takes motion into consideration from two aspects: video depth and optical flow. The former regulates motion by per-frame object distances and spatial layouts, while the later describes motion by cross-frame correspondences that help in preserving fine details and improving temporal consistency. More specifically, given a key frame that exists or generated from text prompts, we first design a diffusion model with spatio-temporal modules to generate the video depth and the corresponding optical flows. Then, the video is generated in the latent space by another spatio-temporal diffusion model under the guidance of depth, optical flow-based warped latent video and the calculated occlusion mask. Lastly, we use optical flows again to align and refine different frames for better video decoding from the latent space to the pixel space. In experiments, MoVideo achieves state-of-the-art results in both text-to-video and image-to-video generation, showing promising prompt consistency, frame consistency and visual quality.

1 Introduction

In video generation, how to generate videos with natural consistent motions is one of the key challenges. In the era of deep learning, Generative Adversarial Networks [21, 50, 56, 61] (GANs) and autoregressive models [18, 31, 59, 63, 64] have become two primary workhorses for video generation due to their great generative modelling abilities. However, GANs are hard to train and may suffer from model collapse, while autoregressive models represent a video as a sequence of tokens from a limited-sized dictionary, which might be insufficient to cover the general video domain.

Recently, diffusion models [52–54] have attracted much attention due to its impressive performance on image generation [29,44,49,76]. They design a forward diffusion process to gradually add noise to the image and a reverse process to gradually remove noise by a learned UNet model [48], under the assumption of the Markov chain with learned Gaussian transitions [28]. To apply diffusion models to video generation, one natural idea is to simply regard the video as a 3D extension of the 2D image and add an extra temporal dimension to the 2D UNet denoising network [16, 25, 27, 30, 51, 75]. However, it might be challenging for the model to learn the video motions implicitly due to the lack of large-scale high-quality video datasets and the limited learning ability of 3D UNet models.

In this paper, we propose to explicitly model and utilize motion for video generation, incorporating depth and optical flow. Depth is employed to guide the per-frame spatial layouts, and a sequence of depth maps is used to capture the movements within the corresponding video. Optical flow, on the other hand, represents correspondences between different frames in the video and can be leveraged for frame alignment, preserving fine details and enhancing temporal consistency. More specifically, as shown in Fig. 1, given an existing key frame

or an image generated by an image latent diffusion model, we first generate the depth and optical flow of the whole video, by utilizing a 3D diffusion model with spatio-temporal modelling blocks. After that, under the joint guidance of depth, optical flow-based warped latent video and the calculated occlusion mask, we generate the video in the latent space with another 3D spatio-temporal diffusion model. The final video is decoded to the pixel space with optical flow-based alignment and feature refinement.

Our contributions are summarized as follows:

1) To the best of our knowledge, we are the first to generate video depth and optical flows from texts or images. We found that a single static image holds clues about the movements of objects and backgrounds, showing great ability in generating video motion.
2) The generated video depth and optical flows are used to jointly control the video motion by regulating the object distances, spatial layouts and cross-frame correspondences. To preserve fine details and improve temporal consistency, different frames are aligned during both latent video generation and decoding stages.
3) Experiments show that our method can generate videos with promising prompt consistency, frame consistency and visual quality, in both text-to-video and image-to-video generation in the open domain.

2 Related Work

Depth and optical flow generation. As two of the fundamental computer vision problems, depth estimation and optical flow estimation have become hot topics for many years [15,36,46] and are widely used in video tasks such as video super-resolution [5, 6], deblurring [38, 39], denoising [7] and frame interpolation [2]. However, to the best of our knowledge, there are nearly no attempts in generating depth maps or optical flows for videos given a conditional input such as text or image. One related method [12] proposes to generate the dense flow map from the sparse flow map input, but it needs strokes by human.

Video generation. Video generation aims to generate videos, mostly under guidance such as text [27,30]. Although GAN-based models [13,22,50,56,61] achieved good results in the past years, most recent video generation models are based either on sequence-to-sequence models [10, 11] or on diffusion models [3, 47, 52–54, 71, 74]. The former line of work first tokenize each video frame into a sequence of discrete tokens and then transform the video generation problem to a sequence-to-sequence translation problem. It could be further divided as autoregressive [18, 31, 64, 68] and non-autoregressive models [60, 69].

The other line of work use the diffusion process for video generation. Ho *et al*. [30] propose a spatial-temporal factorized 3D UNet by adding temporal blocks for video generation, as a natural extension of the standard image diffusion model [52]. Similarly, Ho *et al*. [27] apply the same idea to the cascaded image diffusion model Imagen [49], while He *et al*. [25], Zhou *et al*. [75] and Wang *et*

al. [62] apply it to the latent space [47]. Different from above methods that use spatial-temporal factorized 3D UNet, An *et al*. [1] keep using 2D UNet and enable motion learning by shifting the feature channels along the temporal dimension. Some other methods take the redundancy of videos into consideration. Luo *et al*. [41] represent each frame as the addition of the base frame and residue, while Ge *et al*. [19] encode each frame as the concatenation of the shared latent variable and an individual latent variable.

In particular, training video generation models requires large-scale annotated video data, which are often not publicly available [27,30,60,75]. Therefore, some methods instead propose to generate a video from one of its frames [16,51], as it is fully self-supervised and unlabelled videos are widely available. Singer *et al*. [51] propose a cascaded architecture to generate high-spatiotemporal-resolution videos given an image embedding and a desired frame rate. Esser [16] propose a video editing framework based on the edited image embedding and the depth of the original video. Some other methods try to generate or edit videos based on pre-trained image diffusion models [8, 20, 23, 33, 40, 43, 65, 73]. They often freeze or only optimize some of the model parameters on a single video input. Since our method is proposed for open-domain video generation, the comparison with these zero-shot, one-shot or few-shot methods are omitted here.

Perceptual Video Compression. Generating videos in the latent space is either necessary for sequence-to-sequence models or preferred in diffusion models for the sake of computation burden [47]. Many existing models [25,37,70,75] directly use the pre-trained image VQVAE [58] or VQGAN [17] for encoding and decoding videos, where each frame of the video is processed independently. To improve reconstruction quality, Yu *et al*. [69], He *et al*. [25] and Ruben *et al*. [60] propose several 3D autoencoders to compress the video both spatially and temporally, while Blattmann *et al*. [4] fix the encoder of VQGAN and only add additional temporal layers in the decoder.

3 Method

Due to the lack of large-scale high-quality paired text-video datasets, we limit our setting to uncaptioned video data and try to generate a video from one of its key frames in a fully self-supervised way. To achieve this, we propose a **Mo**tion-aware **Video** generation framework (referred to as MoVideo) with explicit motion modelling. It consists of four stages: key frame generation, video depth and optical flow generation, depth and optical flow-based video generation, and optical flow-augmented video decoding, as shown in Fig. 1. Any public available diffusion model could be used to generate the key frame based on the text prompt or we directly use an existing image as the key frame. The next three stages, including a brief introduction to the diffusion models, are detailed as below.

3.1 Preliminaries on Diffusion Models

Diffusion models [28, 52–54] are probabilistic models that learn a data distribution $p(x)$ by gradually denoising a normally distributed variable $x_T \sim \mathcal{N}(0, 1)$ to obtain the original data $x_0 \sim p(x)$. In the forward diffusion process, we define a fixed Markov chain of length T as

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

where each x_t is obtained by adding a small normally distributed noise to x_{t-1}. α_1 is set to be slightly smaller than 1 and α_t is scheduled to decrease gradually for $t = 1, ..., T$. When T is large enough, *e.g.*, $T = 1000$, x_T is nearly independent of x_0 and meets $x_T \sim \mathcal{N}(0, 1)$.

Since estimating the conditional probability $q(x_{t-1}|x_t)$ is intractable, we approximate it by learning the distribution $p_\theta(x_{t-1}|x_t)$ with parameters θ. Starting from a random noise $x_T \sim \mathcal{N}(0, 1)$, we gradually remove noises with a Markov chain in T steps based on the learned Gaussian transitions. The reverse denoising process is defining as

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

$$p_\theta(x_{0:T}) = p_\theta(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{3}$$

$$p_\theta(x_0) = \int p_\theta(x_{0:T}) dx_{1:T} \tag{4}$$

To optimize the negative log-likelihood of (4), we can derive its variational lower bound by the reparameterization trick [35, 52]. In practice, we empirically use a simplified reweighted variant of variational lower bound [28] as

$$L = \mathbb{E}_{x_0, t\sim[1,T], \epsilon\sim\mathcal{N}(0,1)}\Big[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\Big], \tag{5}$$

where $\epsilon_\theta(\cdot)$ is parameterized by a denoising UNet [48] that takes the noisy sample $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ and the corresponding diffusion timestep t as inputs, and outputs the predicted noise. $\bar{\alpha}_t$ is defined as $\prod_{i=1}^{t} \alpha_i$.

To boost the efficiency, some methods propose to conduct the diffusion process in a compressed latent space defined by a pre-trained autoencoder [17]. When we apply it to the video domain, we can use the image encoder $\mathcal{E}$ to encode a video $x \in \mathcal{R}^{F\times H\times W\times C}$ frame by frame to a latent variable $z = \mathcal{E}(x) \in \mathcal{R}^{F\times h\times w\times c}$, and use the decoder $\mathcal{D}$ to reconstruct the video $\mathcal{D}(z) \approx x$. F, H, W and C are video frame number, height, width and channel number, respectively, in the pixel space, while h, w and c are latent variable height, width and channel number, respectively, in the latent space. In particular, one can further speed up the reverse denoising process by the deterministic sampling [53].

3.2 Video Depth and Optical Flow Generation

Compared with image generation, the main challenge in video generation lies in how to generate motions with good temporal consistency. To describe motions in

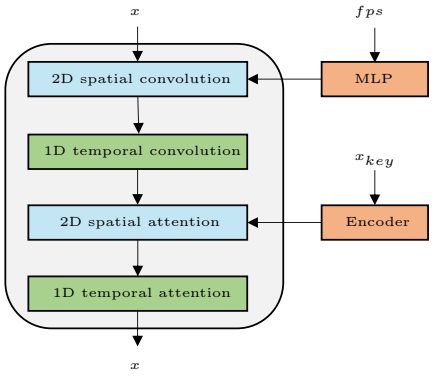

Fig. 2: The basic spatio-temporal block for building the 3D denoising UNet. We add temporal modules, including temporal convolution and temporal attention layers after spatial convolution and spacial attention layers. The fps is encoded by a multi-layer perceptron and then added to the feature after 2D convolution, while key frame x_{key} is encoded by the image encoder from an open-sourced pretrained image-text bi-encoder model and then injected to the 2D spatial attention layer by cross attention.

a video, the most widely used way is calculating the optical flows that represent the correspondences in a video. However, optical flows cannot handle occlusions well and often lead to blurry or deformed motion boundaries [2]. To remedy this, we propose to combine the optical flows with video depth maps that reflect the object distances and spatial layout of each frame, providing accurate information for boundary movements in a video. It is noteworthy that merely using depth maps might be insufficient as they cannot guarantee the temporal consistency of fine details such as textures and colors in different frames. There, before video generation, we jointly generate the video depth and optical flows based on the key frame in this subsection.

Formally, given a center key frame $x_{key} \in \mathcal{R}^{H\times W\times C}$ from a video x, we first use an open-sourced pretrained image-text bi-encoder model $\mathcal{C}$ to extract the image embedding $\mathcal{C}(x_{key})$ before the last average pooling layer. Then, $\mathcal{C}(x_{key})$ is used as the conditional input to control the contents of the generated video depth and optical flows, and the frames per second (fps) is used as an additional condition for controlling the motion magnitude. With these two conditions, we design a diffusion model to learn a joint distribution of depth and optical flows as

$$d, o^{i2v}, o^{v2i} \sim p_\theta\big(d, o^{i2v}, o^{v2i} | \mathcal{C}(x_{key}), fps\big), \tag{6}$$

where $d \in \mathcal{R}^{T\times H\times W\times 1}$ is the video depth that includes the depth maps for each frame, $o^{i2v} \in \mathcal{R}^{T\times H\times W\times 2}$ is the image-to-video optical flow from the key frame to all frames in the video, and $o^{v2i} \in \mathcal{R}^{T\times H\times W\times 2}$ is the video-to-image optical flow. In detail, for the architecture of the denoising UNet, the original 2D network is upgraded to a 3D network that is able to predict the video noise added to the concatenation of video depth and optical flow tensors. As shown in Fig. 2, after each spatial block (including 2D spatial convolution layers and 2D spatial attention layers), we add temporal blocks that consist of several 1D temporal convolution layers and 1D temporal attention layers. The alternative stacking of spatial and temporal blocks allows the network to capture spatio-temporal correlations in videos and generate temporally consistent motions. As for the conditioning mechanism, the image embedding is injected to the model

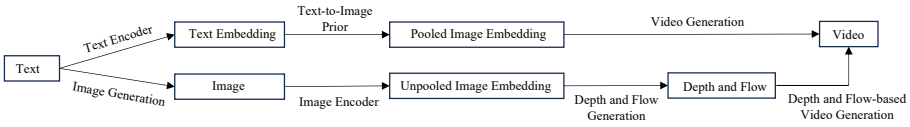

Fig. 3: The comparison on different architectures for text-to-video generation without text-video training pairs. As the top route shows, some methods [16, 51] first encode the text with the text encoder from an open-sourced pretrained image-text bi-encoder model and then use a text-to-image prior [44, 45] to transform it to the pooled image embedding, which is used as the condition to guide the generation of video. Instead, we propose to first generate an image by a public text-to-image latent diffusion model and extract its unpooled image embedding that preserves spatial layout and local details of the image, based on which we generate the depth and optical flow of the video and then use them to guide video generation.

by cross attention [59], after self attention in the 2D spatial attention block. The fps is first transformed to a deep feature by a multi-layer perceptron (MLP) and then added to network after 2D convolution layers.

Remarks. Due to the characteristic difference from RGB video data, there are some remarks on designing a diffusion model for video depth and optical flow generation as below.

1) *Use unpooled image embedding.* Generally, we use outputs of the last average pooling layer of the image encoder as image embeddings [16, 51], as they are aligned with text embeddings and could be obtained by inputting the text embedding to the text-to-image prior [44] during testing, as shown in the top of Fig. 3. However, image semantics, spatial layout and local details might be lost after pooling, posing challenges in generating depth and optical flows. Hence, we use the unpooled image embedding. Additionally, sinusoidal positional encoding [59] is added to the image embedding for encoding spatial information. Nevertheless, above design brings another challenge on training a text-to-image prior for sampling unpooled image embeddings. To tackle with it, as shown in the bottom of Fig. 3, we propose to first generate a RGB image from the text input by a public text-to-image latent diffusion model and then extract its image embedding, allowing us to utilize the strong image diffusion prior to generate the basic semantics of video.

2) *Use image-to-video and video-to-image optical flows.* There are two common choices of describing the video motion with optical flows: flows between neighboring frame pairs or between the key frame and other frames. We use the second design due to following two reasons. First, we can directly obtain all frames from the key frame by optical flow-based warping, avoiding the accumulation of warping errors during propagation. Second, since optical flow only reflects relative movements, our design makes it possible to normalize all flows together, which is found to be critical in flow generation.

3) *Use normalized depth and optical flows.* Since depth maps and optical flows have different numeric ranges, we normalize them separately to be within

-1 and 1 as

$$\widetilde{d} = 2 \times \frac{d - min(d)}{max(d) - min(d)} - 1, \tag{7}$$

$$o_{max} = max\big(max(\|o^{i2v}\|_2), max(\|o^{v2i}\|_2)\big), \tag{8}$$

$$\widetilde{o^{i2v}}, \widetilde{o^{v2i}} = \frac{o^{i2v}}{o_{max}}, \frac{o^{v2i}}{o_{max}}, \tag{9}$$

where $\|\cdot\|_2$ denotes the norm of the optical flow motion vector. However, after normalization, we cannot obtain the original optical flow values and use them for warping, as maximum flow o_{max} is unknown in inference. To remedy this, based on the consistency of video depth maps, we propose an optimization-based method to infer o_{max} from normalized depth maps and optical flows as

$$\begin{aligned} o_{max} = \arg\min_{o_{max}} \sum_{f=1}^{F} \{ & \|\widetilde{d}_f - \mathcal{W}(\widetilde{d}_{key}, \widetilde{o_f^{i2v}} * o_{max})\|_2 + \\ & \|\widetilde{d}_{key} - \mathcal{W}(\widetilde{d}_f, \widetilde{o_f^{v2i}} * o_{max})\|_2 \}, \end{aligned} \tag{10}$$

where $\mathcal{W}$ is the flow-based warping operation. o_{max} is optimized by minimizing the above cost function.

3.3 Depth and Flow-based Video Generation

Given the key frame x_{key}, the frames per second fps, video depth d, image-to-video optical flow o^{i2v} and video-to-image optical flow o^{v2i} generated in subsection 3.2, we first compress the key frame to the latent space as $z_{key} = \mathcal{E}(x_{key})$ with the pre-trained latent encoder [17], and then obtain the occlusion mask m and the warped latent video $\widetilde{z}$ as

$$m_f = \{\|o_f^{i2v} + \mathcal{W}(o_f^{v2i}, o_f^{i2v})\|_2 < \delta_f\}, \tag{11}$$

$$\widetilde{z_f} = \mathcal{W}(z_{key}, o_f^{i2v}) * m_f, \tag{12}$$

where $\delta_f = \alpha(\|o_f^{i2v}\|_2 + \|o_f^{v2i}\|_2) + \beta$ is the threshold for non-occluded regions [42]. α and β are set as 0.01 and 0.5, respectively. $f = 1, ..., F$ represents frame index.

Next, we turn to diffusion models again to learn a conditional distribution of the latent video z as

$$z \sim p_\theta\big(z|\mathcal{C}(x_{key}), fps, d, \widetilde{z}, m\big), \tag{13}$$

where $\mathcal{C}(x_{key})$ is still the image embedding of the key frame as in subsection 3.2. Similarly, we upgrade the 2D UNet to a 3D UNet by inserting temporal layers as shown in Fig. 2. The parameters of the spatial layers are initialized with the original 2D UNet from an open-sourced pretrained latent diffusion model and optimized with one tenth of the global learning rate to preserve the learned image diffusion prior, while the newly added layers are initialized as identical mappings.

The conditioning mechanisms for fps and $\mathcal{C}(x_{key})$ are kept the same as in 3.2. Besides, we concatenate d, $\widetilde{z}$ and m with z as the input to the UNet to guide the video generation process for different purposes. The video depth d controls the video motion and the rough layout of each frame. The warped latent video $\widetilde{z}$ provides semantic information and local details, and it also helps to improve the temporal consistency, especially for texture and color consistencies based on the correspondence information from the optical flow. The occlusion mask m is used to indicate occluded regions, which can tell the network whether the information from $\widetilde{z}$ could be trusted or not.

Particularly, when we concatenate the warped latent video $\widetilde{z}$ as the input condition, we empirically found that it leads to unsatisfactory motions in the generated video, possibly because the reconstruction process is sometimes misled by the imperfect motions from $\widetilde{z}$. To alleviate the problem, we randomly mask $\widetilde{z}$ as all-zeros with a probability of 0.5, which prevents the network from learning bad motions of $\widetilde{z}$ but still allows it to benefit from the semantic information and local details of $\widetilde{z}$.

3.4 Optical Flow-Augmented Video Decoding

After obtaining the latent video z from the last subsection 3.3, the final step is to decode it to the pixel space as video x. Since different video frames are highly related but misaligned, we propose to align different frames for joint decoding. As shown in Fig. 4, we upgrade the 2D decoder in [17] to a 3D decoder by inserting temporal convolution layers after spatial convolution layers. Then, we fuse cross-frame information by explicitly aligning the key frame feature z_{key} towards the each frame feature z_f with flow-guided deformable convolution [9, 14] as

$$z'_f = \mathcal{W}(z_{ref}, o_f^{i2v}), \tag{14}$$

$$s_f, a_f = CNN(Concat(o_f^{i2v}, z_f{}', z_f)), \tag{15}$$

$$\widehat{z_f} = DC(z_{ref}, o_f^{i2v} + s_f, a_f), \tag{16}$$

where the offset s_f and mask a_f are estimated according the concatenation of optical flow o_f^{i2v}, the warped feature $z_f{}'$ and the frame z_f. CNN is a neural network with several 3×3 convolution layers and DC is the deformable convolution. Next, we concatenate the aligned feature $\widehat{z_f}$ with z_f and use several ResNet blocks [24] (denoted as $\mathcal{R}$) to refine the feature as

$$z_f = z_f + \mathcal{R}(Concat(\widehat{z_f}, z_f)), \tag{17}$$

which is used for each stage of the decoder. Hence, the decoding of every frame f is aware of other frames and especially obtains information from the key frame.

In training, we use a combination of pixel loss, perceptual loss [32] and GAN loss [22] following VQGAN [17]. To improve temporal coherence, we further extract features from S3D [66] for 3D perceptual loss and upgrade the 2D discriminator to a 3D discriminator for the 3D GAN loss by replacing the 2D

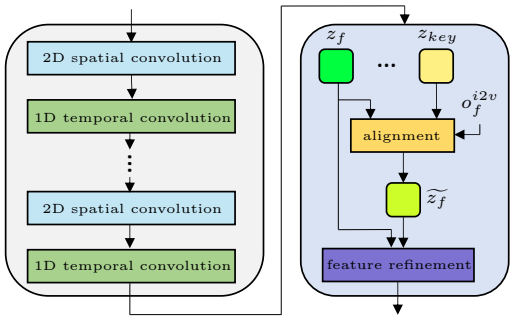

Fig. 4: The basic block for building the optical flow-augmented video decoding model. We add temporal convolution layers after spatial convolution layers to extract spatio-temporal video features. After that, with the optical flow o_f^{i2v}, we align the key frame z_{key} towards the each frame as $\widetilde{z_f}$, which is concatenated with z_f for feature refinement.

convolutions with 3D convolutions. We train the decoder with a combination of losses between the decoded video $\hat{x}$ and ground-truth video x as:

$$\begin{aligned} L = L_1(\hat{x}, x) + \lambda_1 \sum_{f=1}^{F} L_{percep\_2d}(\hat{x}, x) + \lambda_2 L_{percep\_3d}(\hat{x}, x) \\ + \lambda_3 \sum_{f=1}^{F} L_{GAN\_2d}(\hat{x}, x) + \lambda_4 L_{GAN\_3d}(\hat{x}, x). \end{aligned} \tag{18}$$

4 Experiments

4.1 Experimental Setup

Architecture. For video generation, we use the spatial blocks from the 2D UNet of an open-sourced pretrained latent diffusion model and add the same numbers of temporal blocks with similar channel numbers to upgrade it to a 3D UNet. For video depth and optical flow generation, we reduce the channel numbers by half as depth maps and optical flows have less details. For video decoding, we upgrade the 2D VQGAN decoder [17] to a 3D decoder in a similar way. The details on the networks are provided in the supplementary.

Training and testing. In training, we use the similar training set as in [51] and randomly cut 17-frame video clips with a random fps between 3 and 30. Each video clip is resized to ensure its shorter side measures 256 pixels, followed by a center-crop of size 256×256. The depth maps and optical flows are estimated by open-sourced pretrained depth and optical flow models. The depth and optical flow generation model, video generation model (includes the fine-tuning stage) and video decoding model are trained separately with different training iterations, batch sizes and optimizer settings. In testing, we generate $17 \times 256 \times 256$ videos at a fps of 3. More details are summarized in the supplementary.

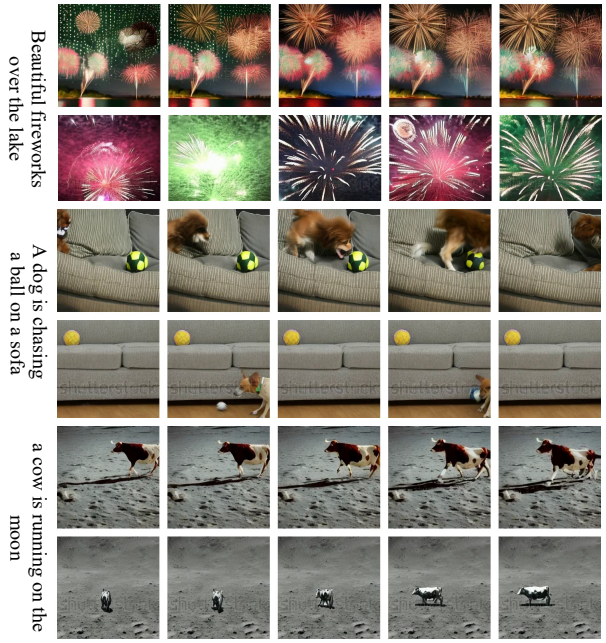

Fig. 5: Visual comparison on text-to-video generation. For each example, the first row is from our method, while the second row is from VideoDiffusion [70]. More visual comparisons, including video results, are provided in the supplementary.

4.2 Text-to-Video Generation

We test the text-to-video generation ability of our model on multiple text prompts in Figures 5. We mainly compare our method with one of the recent best models VideoFusion [41], as it is the only open-sourced text-to-video model during the time when we did experiments (to the best of our knowledge) and it uses the same publicly available training set as in our model. As we can see, our model can generate high-quality visually-pleasing videos with good consistency in three aspects. First, it is consistent with the semantics of text prompts, including the objects (*e.g.*, "fireworks" and "lake" in the first example) and the motion (*e.g.*, "chasing" in the second example). In contrast, VideoFusion fails to generate the "lake" as the background or show the "chasing" action between the "dog" and "ball". Second, across different frames, the generated objects of our method are consistent (*e.g.*, the same "fireworks" in the first example), while VideoFusion generates "fireworks" with different colors and shapes for neighbouring frames. Third, our method is able to generate natural consistent motions without sudden large movements, *e.g.*, the "chasing" of "dog" is natural), but the "dog" generated by VideoFusion may disappear suddenly in some frames. The visualization of intermediate results and more comparisons are provided in the supplementary due to page limit.

For quantitative comparison, we compare our method on zero-shot text-to-video generation on UCF-101 [55] and MSR-VTT [67]. As shown in Tables 1 and 2, our method achieves best performance on all metrics. Besides, we choose 100 text prompts to generate videos in the open domain. As shown in Table 3,

Table 1: Quantitative comparison of zero-shot text-to-video generation on UCF-101 [55].

Method	IS↑	FVD↓
CogVideo [31]	25.27	701.59
Make-A-Video [51]	33.00	367.23
Video LDM [4]	33.45	550.61
MoVideo (ours)	34.13	313.41

Table 2: Quantitative comparison of zero-shot text-to-video generation on MSR-VTT [67].

Method	FID↓	CLIPSIM↑
NÜWA [64]	47.68	0.2439
CogVideo [31]	23.59	0.2631
Latent-Shift [1]	15.23	0.2773
Make-A-Video [51]	13.17	0.3049
Video LDM [4]	-	0.2929
MoVideo (ours)	12.71	0.3213

Table 3: Quantitative comparison of open-domain text-to-video generation.

Method	Frame Consistency↑	Prompt Consistency↑	Preference Rate↑
VideoDiffusion [41]	0.9759	0.3143	18.7%
MoVideo (ours)	0.9867	0.3631	81.3%

Table 4: Quantitative comparison of image-to-video generation on DAVIS [34].

Method	PSNR↑	LPIPS↓	FID↓	FVD↓
Gen1 [16]	28.39	0.290	149.73	973.71
MoVideo (ours)	29.88	0.079	40.14	335.53

generated videos of our method is more consistent with the prompts and across the frames. It is preferred by 81.3% of the 30 users on average during user study.

4.3 Image-to-Video Generation

We show the results for image-to-video generation in Fig. 6 on the DAVIS [34] dataset. We use the flow-based warping as the baseline and mainly compare our method with the recent best method Gen1 [16], which is re-implemented by us and trained with the same training settings for fair comparison. As shown in the figure, the generated video of our method is almost the same as the ground-truth video, while the Gen1 changes the background and object appearance. The quantitative results in Table 4 further validate our observation from the aspects of fidelity (reflected by PSNR) and perceptual quality (reflected by LPIPS [72], FID [26] and FVD [57]). When we use the generated depth and optical flows from the image, our method is able to generate a video with similar semantics but different motions.

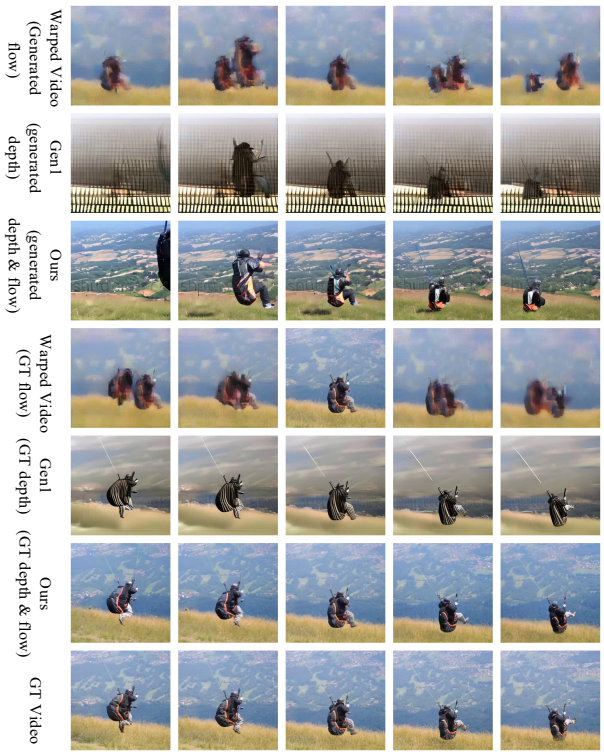

Fig. 6: Visual comparison on image-to-video generation. The first three rows are guided by the generated depth and optical flows, while the rest rows are guided by the ground-truth (GT) ones. More visual comparisons, including video results, are provided in the supplementary.

4.4 Ablation Studies

Ablation study on the warped video. We obtain a warped video by warping the key frame with the image-to-video optical flow, and then use it to guide the video generation. As shown in Table 6, there exists a large margin between with and without warped video, which indicates that the concatenation of the warped video plays a critical role in the model performance. In fact, without warped video, the model cannot preserve the textures and colors. A visual example is provided in the supplementary.

Ablation study on the occlusion mask. The occlusion mask is concatenated as the conditional input to help the model identify the occluded regions. Without the occlusion mask, the performance drops as show in Table 6, possibly due to the wrong correspondences appearing in the occluded regions of videos. In this case, we found that the ghosting artifact often appears, as provided in the supplementary.

Ablation study on flow-augmented video decoding. The optical flow is used to align different frames during decoding. To validate its effectiveness, we encode the videos from DAVIS [34] with the pre-trained image encoder [17], and use different decoders to decode the video. As shown in Table 5, when upgrading the 2D decoder to a 3D one, the perceptual metrics become better with slight

Table 5: Ablation study on video decoding.

Decoder	2D decoder	3D decoder	MoVideo (ours)
PSNR↑	30.11	29.99	29.88
LPIPS↓	0.093	0.087	0.079
FID↓	43.28	42.71	40.14
FVD↓	401.66	379.73	335.53

Table 6: Ablation study on warped video and occlusion mask.

Warped Video	Occlusion Mask	PSNR↑	LPIPS↓	FID↓	FVD↓
	✓	27.92	0.303	153.73	796.82
✓		29.45	0.093	45.61	343.90
✓	✓	29.88	0.079	40.14	335.53

drop in PSNR. When using the proposed flow-augmented video decoder, we achieve the best results on all perceptual metrics, indicating its ability to decode better visually-pleasing videos.

To to page limit, the accompanying qualitative results, video results, as well as more ablation studies are provided in the supplementary.

5 Conclusion

In this paper, we proposed a motion-aware video generation framework (MoVideo) that consists of four stages: key frame generation, video depth and optical flow generation, depth and optical flow-based video generation, and optical flow-augmented video decoding. Based on an text-to-image generation model, we use the text prompt to generate an image as the key frame, which is used to guide the generation of video motion represented by video depth, image-to-video optical flow and video-to-image optical flow. These motion representations could be further utilized to guide the latent video generation and assist the video decoding processes. Experiments demonstrate that our method achieves state-of-the-art performance on both text-to-video and image-to-video generation.

Limitation and potential negative impact Although our multi-stage design allows explicit motion modelling, control and guidance, it suffers from multiple training and inference steps. Besides, the performance is highly related to optical flow and depth estimation methods as we use predicted pseudo-labels in training. These limitations could be our future working directions. For potential negative impact, this model might have the risks of data breaches or misinformation.

Acknowledgements

Jingyun Liang and Luc Van Gool were partially supported by the ETH Zurich Fund (OK), a Huawei Technologies Oy (Finland) project and the China Scholarship Council. Radu Timofte was partially supported by The Alexander von Humboldt Foundation. Yuchen Fan and Rakesh Ranjan from Meta did not receive above fundings.

References

1. An, J., Zhang, S., Yang, H., Gupta, S., Huang, J.B., Luo, J., Yin, X.: Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. arXiv preprint arXiv:2304.08477 (2023)
2. Bao, W., Lai, W.S., Ma, C., Zhang, X., Gao, Z., Yang, M.H.: Depth-aware video frame interpolation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3703–3712 (2019)
3. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Li, Y., Michaeli, T., et al.: Lumiere: A space-time diffusion model for video generation. arXiv preprint arXiv:2401.12945 (2024)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
5. Cao, J., Li, Y., Zhang, K., Liang, J., Van Gool, L.: Video super-resolution transformer. arXiv preprint arXiv:2106.06847 (2021)
6. Cao, J., Liang, J., Zhang, K., Wang, W., Wang, Q., Zhang, Y., Tang, H., Van Gool, L.: Towards interpretable video super-resolution via alternating optimization. In: European Conference on Computer Vision. pp. 393–411 (2022)
7. Cao, J., Wang, Q., Liang, J., Zhang, Y., Zhang, K., Van Gool, L.: Learning task-oriented flows to mutually guide feature alignment in synthesized and real video denoising. arXiv preprint arXiv:2208.11803 (2022)
8. Ceylan, D., Huang, C.H.P., Mitra, N.J.: Pix2video: Video editing using image diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23206–23217 (2023)
9. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. arXiv preprint arXiv:2104.13371 (2021)
10. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
11. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11315–11325 (2022)
12. Chen, T.S., Lin, C.H., Tseng, H.Y., Lin, T.Y., Yang, M.H.: Motion-conditioned diffusion model for controllable video synthesis. arXiv preprint arXiv:2304.14404 (2023)
13. Clark, A., Donahue, J., Simonyan, K.: Adversarial video generation on complex datasets. arXiv preprint arXiv:1907.06571 (2019)

14. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: IEEE International Conference on Computer Vision. pp. 764–773 (2017)
15. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: IEEE International Conference on Computer Vision. pp. 2758–2766 (2015)
16. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7346–7356 (2023)
17. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
18. Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.B., Parikh, D.: Long video generation with time-agnostic vqgan and time-sensitive transformer. In: European Conference on Computer Vision. pp. 102–118. Springer (2022)
19. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22930–22941 (2023)
20. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)
21. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. pp. 2672–2680 (2014)
22. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
23. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
25. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221 (2022)
26. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
27. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
28. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
29. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. The Journal of Machine Learning Research **23**(1), 2249–2281 (2022)
30. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv preprint arXiv:2204.03458 (2022)

31. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
32. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. pp. 694–711. Springer (2016)
33. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023)
34. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: Asian Conference on Computer Vision. pp. 123–141 (2018)
35. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
36. Li, R., Gong, D., Yin, W., Chen, H., Zhu, Y., Wang, K., Chen, X., Sun, J., Zhang, Y.: Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21539–21548 (2023)
37. Li, X., Chu, W., Wu, Y., Yuan, W., Liu, F., Zhang, Q., Li, F., Feng, H., Ding, E., Wang, J.: Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. arXiv preprint arXiv:2309.00398 (2023)
38. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: VRT: A video restoration transformer. IEEE Transactions on Image Processing **33**, 2171–2182 (2024). https://doi.org/10.1109/TIP.2024.3372454
39. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Van Gool, L.: Recurrent video restoration transformer with guided deformable attention. In: Advances in Neural Information Processing Systems. pp. 378–393 (2022)
40. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. arXiv preprint arXiv:2303.04761 (2023)
41. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10209–10218 (2023)
42. Meister, S., Hur, J., Roth, S.: Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
43. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. arXiv preprint arXiv:2303.09535 (2023)
44. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
45. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
46. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)

47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
48. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241 (2015)
49. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
50. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: Proceedings of the IEEE international conference on computer vision. pp. 2830–2839 (2017)
51. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
52. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015)
53. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
54. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
55. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
56. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1526–1535 (2018)
57. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
58. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
60. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022)
61. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. Advances in neural information processing systems **29** (2016)
62. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
63. Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., Duan, N.: Godiva: Generating open-domain videos from natural descriptions. arXiv preprint arXiv:2104.14806 (2021)
64. Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., Duan, N.: Nüwa: Visual synthesis pre-training for neural visual world creation. In: European conference on computer vision. pp. 720–736. Springer (2022)

65. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
66. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 305–321 (2018)
67. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016)
68. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
69. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: Magvit: Masked generative video transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10459–10469 (2023)
70. Yu, S., Sohn, K., Kim, S., Shin, J.: Video probabilistic diffusion models in projected latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18456–18466 (2023)
71. Zhang, J., Li, X., Wan, Z., Wang, C., Liao, J.: Text2nerf: Text-driven 3d scene generation with neural radiance fields. arXiv preprint arXiv:2305.11588 (2023)
72. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
73. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)
74. Zhang, Y., Wei, Y., Lin, X., Hui, Z., Ren, P., Xie, X., Ji, X., Zuo, W.: Videoelevator: Elevating video generation quality with versatile text-to-image diffusion models. arXiv preprint arXiv:2403.05438 (2024)
75. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)
76. Zhu, Y., Zhang, K., Liang, J., Cao, J., Wen, B., Timofte, R., Van Gool, L.: Denoising diffusion models for plug-and-play image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1219–1229 (2023)