MotionDirector: Motion Customization of Text-to-Video Diffusion Models

Rui Zhao[1], Yuchao Gu[1], Jay Zhangjie Wu[1], David Junhao Zhang[1], Jia-Wei Liu[1], Weijia Wu[1], Jussi Keppo[2], and Mike Zheng Shou[1] *

[1]Show Lab, [2]National University of Singapore, Singapore

Abstract. Large-scale pre-trained diffusion models have exhibited remarkable capabilities in diverse video generations. Given a set of video clips of the same motion concept, the task of Motion Customization is to adapt existing text-to-video diffusion models to generate videos with this motion. Adaptation methods have been developed for customizing appearance like subject or style, yet under-explored for motion. It is straightforward to extend mainstream adaption methods for motion customization, including full model tuning and Low-Rank Adaptions (LoRAs). However, the motion concept learned by these methods is often coupled with the limited appearances in the training videos, making it difficult to generalize the customized motion to other appearances. To overcome this challenge, we propose MotionDirector, with a dual-path LoRAs architecture to decouple the learning of appearance and motion. Further, we design a novel appearance-debiased temporal loss to mitigate the influence of appearance on the temporal training objective. Experimental results show the proposed method can generate videos of diverse appearances for the customized motions. Our method also supports various downstream applications, such as the mixing of different videos with their appearance and motion respectively, and animating a single image with customized motions. The project website is at: MotionDirector.

Keywords: motion customization · text-to-video generation · diffusion models

1 Introduction

Text-to-video diffusion models [17,21,50] are approaching generating high-quality diverse videos given text instructions. The open-sourcing of foundational text-to-video models [55, 60] pre-trained on large-scale data has sparked enthusiasm for video generation in both the community and academia. Users can create videos that are either realistic or imaginative simply by providing text prompts. While foundation models generate diverse videos from the same text, adapting them to generate more specific content can better accommodate the preferences of users. Similar to the customization of text-to-image foundation models [45], tuning video foundation models to generate videos of a certain concept of appearance,

* Corresponding Author.

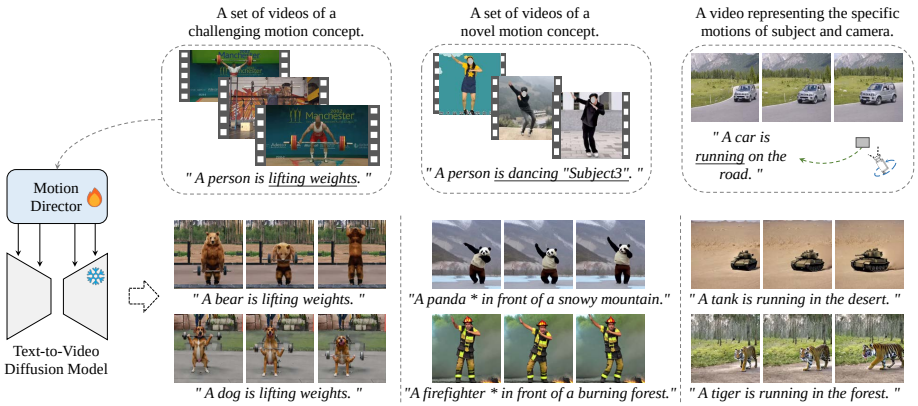

Fig. 1: Motion customization of the text-to-video diffusion.

like subject or style, has also been explored [17]. Compared with images, videos consist of not just appearances but also motion dynamics. Users may desire to create videos with specific motions, such as a car moving forward and then turning left under a predefined camera perspective, as illustrated on the right side in Fig. 1. However, customizing the motions in the text-to-video generation is still under-explored.

The task of Motion Customization is formulated as follows: given a set of reference videos representing a specific motion concept, the objective is to customize pre-trained foundational models to generate videos that accurately capture and replicate the depicted motion. The customized motion concept can be challenging motions that are hard for the pre-trained model to generate, novel motions not present in the pre-training dataset, and specific motions that are difficult to express in text. Examples of these are respectively showcased in the first, second, and third columns of Fig 1. In contrast, previous works on appearance customization adapt the foundation models to generate samples with desired appearance, like subject or style, given reference videos or images representing such appearance [17, 45]. It is straightforward to use previous adaption methods for motion customization. For example, on the given reference videos, fine-tuning the weights of foundation models [45], parameter-efficient tuning additional layers [70], or training Low-Rank Adaptions (LoRAs) [26] injected in the layers of foundation models. However, customizing diffusion models to generate desired motions without harming their appearance diversity is challenging because the motion and appearance are coupled with each other at the step-by-step denoising stage. Directly deploying previous adaption methods to learn motions makes the models fit the limited appearances seen in the reference videos, posing challenges in generalizing the learned motions to various appearances. Recent works on controllable text-to-video generations [11, 17, 63] generate videos controlled by signals representing pre-defined motions. However, the control signals, such as depth maps or edges, impose constraints on the shapes of subjects and back-

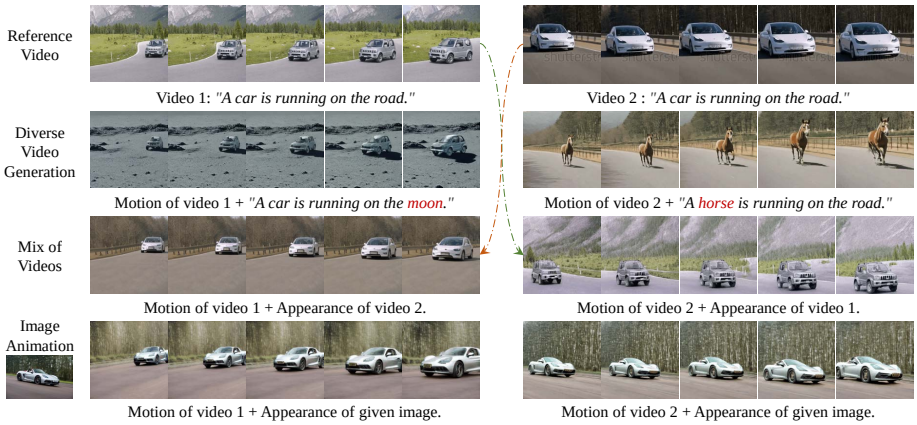

Fig. 2: (Row 1) Take two videos to train the proposed MotionDirector, respectively. (Row 2) MotionDirector can generalize the learned motions to diverse appearances. (Row 3) MotionDirector can mix the learned motion and appearance from different videos to generate new videos. (Row 4) MotionDirector can animate a single image with learned motions.

grounds, thus influencing the appearance of generated videos in a coupled way. Besides, these methods accept only one sequence of control signals to generate one video, which may not be suitable for users seeking certain motion types without strict spatial constraints, such as the example of lifting weights in Fig. 1.

To achieve motion customization of text-to-video diffusion models while preserving appearance diversity, we propose the MotionDirector, which tunes the foundation models to learn the appearance and motions in the given single or multiple reference videos in a decoupled way. MotionDirector tunes the models with low-rank adaptions (LoRAs) while keeping their pre-trained parameters fixed to retain the learned generation knowledge. Specifically, the MotionDirector employs a dual-path architecture, as shown in Fig. 3. For each video, a spatial path consists of a foundation model with trainable spatial LoRAs injected into its spatial transformer layers. These spatial LoRAs are trained on a single frame randomly sampled per training step to capture the appearance characteristics of the input videos. The temporal path, on the other hand, is a replica of the foundation model that shares the spatial LoRAs with the spatial path to fit the appearance of the corresponding input video. Additionally, the temporal transformers in this path are equipped with temporal LoRAs, which are trained on multiple frames of input videos to capture the underlying motion patterns. To further enhance the learning of motions, we propose an appearance-debiased temporal loss to mitigate the influence of appearance on the temporal training objective.

Only deploying the trained temporal LoRAs enables the foundation model to generate videos of the learned motions with diverse appearances, as shown in the second row of Fig 2. The decoupled paradigm further makes an interesting

kind of video generation feasible, which is the mix of the appearance from one video with the motion from another video, called the mix of videos, as shown in the third row of Fig 2. The key to this success lies in that MotionDirector can decouple the appearance and motion of videos and then combine them from various source videos. It is achieved by injecting spatial LoRAs trained on one video and temporal LoRAs trained on another video into the foundation model. Besides, the learned motions can be deployed to animate images, as images can be treated as appearance providers, as shown in the last row of Fig 2.

We conducted experiments on two benchmarks with 86 different motions and over 600 text prompts to test proposed methods, baselines, and comparison methods. The results show our method can be applied to different diffusion-based foundation models and achieve motion customization of various motion concepts. Compared with other methods, our method avoids fitting the limited appearance of reference videos, and can generalize the learned motions to diverse appearances.

Our contributions are summarized as follows:

- We introduce and define the task of Motion Customization. The challenge lies in generalizing the customized motions to various appearances.
- We propose the MotionDirector with a dual-path architecture and a novel appearance-debiased temporal training objective, to decouple the learning of appearance and motion.
- Experiments on two benchmarks demonstrate that MotionDirector can customize various base models to generate diverse videos with desired motion concepts, and outperforms controllable generation methods and tuning-based methods.

2 Related Work

2.1 Text-to-Video Generation.

To achieve high-quality video generation, various methods have been developed, such as Generative Adversarial Networks (GANs) [2, 47, 49, 56, 57, 59], autoregressive models [12, 25, 33, 54, 75] and implicit neural representations [51, 80]. Diffusion-based models [3, 10, 13, 18, 22–24, 27, 34, 35, 40, 41, 50, 58, 60, 62, 65, 71, 78, 79, 81, 84, 86] are also approaching high-quality generation by training conditional 3D U-Nets [44] to denoise from randomly sampled sequences of Gaussian noises [74]. Recent foundation models [4, 17, 21, 36, 50, 64, 82] are pre-trained on large-scale image and video datasets [1, 8, 48], to learn powerful generation ability. Some works turn text-to-image foundation models to text-to-video generation by manipulation on cross-frame attention or training additional temporal layers, like Tune-A-Video [71], Text2Video-Zero [30], and AnimiteDiff [15]. The recently open-sourced foundation models [55, 60] have ignited enthusiasm among users to generate realistic or imaginative videos, and further make it possible for users to customize and build their own private models.

2.2 Generation Model Customization.

Customizing the pre-trained large foundation models can fit the preferences of users better while maintaining powerful generation knowledge without training from scratch. Previous customization methods for text-to-image diffusion models [7,14,32,45,52,68] aim to generate certain subjects or styles, given a set of example images. Dreambooth [45] or LoRA [26] can be simply applied to customizing video foundation models to generate videos with certain subjects or styles, given a set of reference video clips or images. The recently proposed VideoCrafter [19] has explored this, which we categorize as appearance customization. In addition to appearances, videos are also characterized by the motion dynamics of subjects and camera movements across frames. After we proposed the task of motion customization, some concurrent works [28, 39, 43, 66, 67, 72, 76] emerged in this field. DreamVideo [67] trains several adapters in U-Net to customize both the subject and motion in video generation. Lamp [72] learns a motion pattern from a set of videos to utilize image diffusion models to generate videos.

2.3 Controllable Video Generation.

Controllable generation aims to ensure the generation results align with the given explicit control signals, such as depth maps, human pose, optical flows, etc. [5, 16, 29, 37, 38, 42, 61, 73, 77, 83, 85]. For the controllable text-to-video generation methods, i.e. the VideoCrafter [17], VideoComposer [63], Control-A-Video [6], they train additional branches that take condition signals to align the generated videos with them. Unlike the human poses for specifically controlling the generation of human bodies, the general control singles, such as depth maps, are typically extracted from reference videos and are coupled with both appearance and motion. This results in the generation results being influenced by both the appearance and motion in reference videos. Applying these methods directly in motion customization is challenging when it comes to generalizing the desired motions to diverse appearances.

3 Methodology

3.1 Preliminaries

Video Diffusion Model. Video diffusion models train a 3D U-Net to denoise from a randomly sampled sequence of Gaussian noises to generate videos, guided by text prompts. The 3D U-net basically consists of down-sample, middle, and up-sample blocks. Each block has several convolution layers, spatial transformers, and temporal transformers as shown in Fig 3. The 3D U-Net ϵ_θ and a text encoder τ_θ are jointly optimized by the noise-prediction loss, as detailed in [9]:

$$\mathcal{L} = \mathbb{E}_{z_0, y, \epsilon \sim \mathcal{N}(0,I), t \sim \mathcal{U}(0,T)} \left[\| \epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y)) \|_2^2 \right], \tag{1}$$

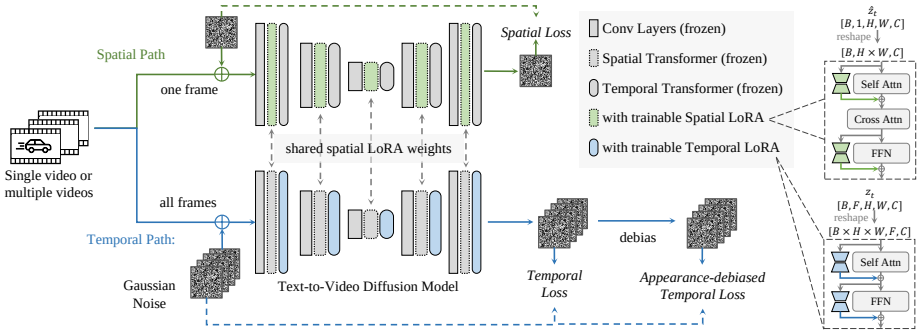

Fig. 3: The dual-path architecture of the proposed method. All pre-trained weights of the base diffusion model remain fixed. In the spatial path, the spatial transformers are injected with trainable spatial LoRAs as shown on the right side. In the temporal path, the spatial transformers are injected with spatial LoRAs sharing weights with those ones in the spatial path, and the temporal transformers are injected with trainable temporal LoRAs.

where z_0 is the latent code of the training videos, y is the text prompt, ϵ is the Gaussian noise added to the latent code, and t is the time step. As discussed in [9], the noised latent code z_t is determined as:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \ \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_t, \tag{2}$$

where α_t is a hyper-parameter controlling the noise strength.

Low-Rank Adaption. Low-rank adaption (LoRA) [26] was proposed to adapt the pre-trained large language models to downstream tasks. Recently it has been applied in text-to-image generation and text-to-video generation tasks to achieve appearance customization [19, 46]. LoRA employs a low-rank factorization technique to update the weight matrix W as

$$W = W_0 + \Delta W = W_0 + BA, \tag{3}$$

where $W_0 \in \mathbb{R}^{d\times k}$ represents the original weights of the pre-trained model, $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ represent the low-rank factors, where r is much smaller than original dimensions d and k. LoRA requires smaller computing sources than fine-tuning the weights of the entire network like DreamBooth [45], and it is convenient to spread and deploy as a plug-and-play plugin for pre-trained models.

In the following sections, we introduce the proposed MotionDirector, containing the dual-path Low-Rank Adaptions (in Sec. 3.2) for learning the appearance and motion in reference videos in a decoupled manner, and appearance-debiased temporal loss (in Sec. 3.3)) to further mitigate the influence of appearance on the temporal training objective.

Fig. 4: Comparison of videos generated by the base model, the model fine-tuned in the coupled manner, the model fine-tuned without the proposed Appearance-Debiased Temporal Loss (AD-Loss), and our full-version MotionDirector method.

3.2 Dual-Path Low-Rank Adaptions

At each time-step t, the 3D U-Net takes in the latent code $z_t \in \mathbb{R}^{b\times f\times w\times h\times c}$ and the conditional input y (e.g., text), where b, f, w, h, c represents the size of the batch, frame, width, height, and channel dimensions, respectively. The spatial transformers apply spatial self-attention along the spatial dimensions w, h to improve the correlation between pixels, and then leverage the cross-attention between the latent code and the conditional input y to improve textual alignment. The temporal transformers apply temporal self-attention along the frame dimension f to improve the temporal consistency between frames. However, spatial and temporal information in the latent code gradually become coupled with each other during the step-by-step denoising stage. Attempting to directly learn and fit the motions in reference videos will inevitably lead to fitting their limited appearances. As shown in the third row "Coupled Tuned" of Fig. 4. From the results on the left side, we can observe that the generated "panda" is coupled with the appearance of "human" in the reference videos and the generated background is also coupled with that of reference videos. Besides, the results on the right side show that the coupled tuned model can not generate the "Eiffel Tower" because it is overfitted on the limited backgrounds in the reference videos.

To address this problem, we propose to tune the spatial and temporal transformers in a dual-path way to learn the earn the appearance and motion in reference videos, respectively, as shown in Fig. 3. Specifically, for the spatial path, we inject LoRAs into spatial transformers to learn the appearance of training data, and for the temporal path, we inject LoRAs into temporal transformers to learn the motion in videos. The training of spatial and temporal LoRAs are as follows.

Spatial Training. For the spatial path, we inject unique spatial LoRAs into the spatial transformers for each training video while keeping the weights of pre-trained 3D U-Net fixed. To maintain the learned strong and diverse textual

alignment ability, we do not inject LoRAs into cross-attention layers of spatial transformers, since their weights influence the correlations between the pixels and text prompts. On the other hand, we inject LoRAs into spatial self-attention layers and feed-forward layers to update the correlations in spatial dimensions to enable the model to reconstruct the appearance of training data. For each training step, the spatial LoRAs are trained on a single frame randomly sampled from the training video to fit its appearance while ignoring its motion, based on spatial loss, which is reformulated as

$$\mathcal{L}_{spatial} = \mathbb{E}_{z_0,y,\epsilon,t,i\sim\mathcal{U}(0,F)}\left[\|\epsilon - \epsilon_\theta(z_{t,i}, t, \tau_\theta(y))\|_2^2\right], \tag{4}$$

where F is the number of frames of the training data and the $z_{t,i}$ is the sampled frame from the latent code z_t.

Temporal Training. For the temporal path, we inject the temporal LoRAs into self-attention and feed-forward layers of temporal transformers to update the correlations along the frame dimension. Besides, the spatial transformers are injected with LoRAs sharing the same weights learned from the spatial path, to force the trainable temporal LoRAs to ignore the appearance of the training data. The temporal LoRAs could be simply trained on all frames of training data based on the temporal loss $\mathcal{L}_{org\text{-}temp}$, formulated in the same way as equation (1).

3.3 Appearance-Debiased Temporal Loss

The videos generated by models trained with original temporal loss are shown in the fourth row of Fig. 4. We can observe that the motions are not learned well. To explore the underlying reasons, we visualized the denoising process of several videos.

As illustrated in Fig. 5, when considering the latent codes of each frame ${z_{t,i}}_{i=1}^{F}$ as a set of points in the latent space, we found that motion primarily impacts the underlying dependencies between these point sets, whereas the distances between different sets of points are more influenced by appearance. The original temporal training objective does not take this into account. The learning of motion dynamics, embedded within the internal connectivity structure of latent codes, remains biased by positional differences of latent codes. These biases are primarily influenced by the appearance of training samples and pose a challenge to learning motions efficiently.

To further decouple the motion from appearance, we proposed to eliminate the appearance bias among the noises and predicted noises, and calculate the appearance-debiased temporal loss on them. The debiasing of each noise $\epsilon_i \in \{\epsilon_i\}_{i=1}^{F}$ is as follows,

$$\phi(\epsilon_i) = \sqrt{\beta^2 + 1}\epsilon_i - \beta\epsilon_{anchor}, \tag{5}$$

where β is the strength factor controlling the debiasing strength and a higher value leads to stronger debiasing effects. The ϵ_{anchor} is the anchor among the

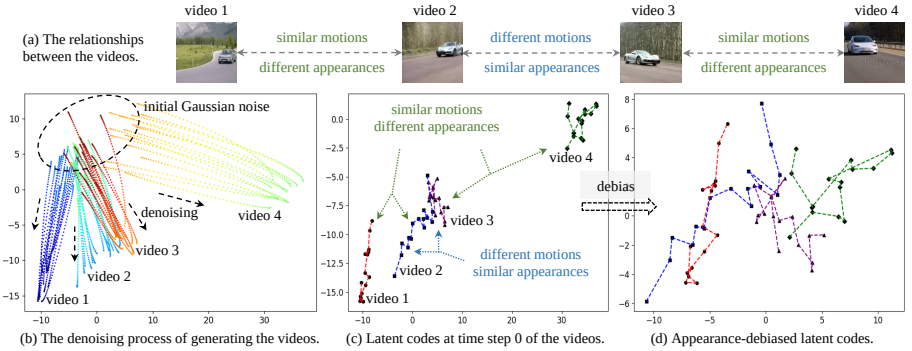

Fig. 5: (a) Four example videos (the same as the videos in the first and fourth rows of Fig. 2) and their relationships in terms of motion and appearance. (b) We inverse the four videos based on the video diffusion model and visualize the denoising process. Each point corresponds to a latent code $z_{t,i,j}$ at time step t of i-th frame of j-th video, reduced to 2 dimensions through PCA. (c) Take latent codes at time step 0 for example, the ones of the same video are connected in order of frames. We find that the internal connectivity structure between latent codes is more influenced by motion, while the distance between sets of latent codes is primarily affected by the difference in appearance. (d) Through the proposed appearance-debiased operation, as formulated in Equation. 5, the latent codes eliminate distance bias introduced by appearance, while preserving the internal connectivity structure that represents motion dynamics.

frames from the same training data. If β equals 1, it indicates that we attribute equal importance to both appearance and motion in the video. The detailed mathematical derivation of this equation is provided in the supplementary materials. In practice, we find that simply setting $\beta = 1$ and randomly sampling $\epsilon_i \in \{\epsilon_i\}_{i=1}^{F}$ as the anchor can achieve good performance. The appearance-debiased temporal loss (AD-Loss) is reformulated as

$$\mathcal{L}_{ad\text{-}temp} = \mathbb{E}_{z_0,y,\epsilon,t}\left[\|\phi(\epsilon) - \phi(\epsilon_\theta(z_t, t, \tau_\theta(y)))\|_2^2\right]. \tag{6}$$

For temporal LoRAs, the loss function is the combination of temporal loss and appearance-debiased temporal loss as follows,

$$\mathcal{L}_{temporal} = \mathcal{L}_{org\text{-}temp} + \mathcal{L}_{ad\text{-}temp}. \tag{7}$$

3.4 Inference of MotionDirector

In the inference stage, we inject the trained temporal LoRAs into the pre-trained video diffusion model to enable it to generate diverse videos with the learned motion from the training data. If the training data is a single video, the learned motion will be a specific motion, such as an object first moving forward and then turning to the left. If the training data is a set of videos, the learned motion will be the motion concept provided by them, like lifting weights or playing golf.

Fig. 6: Qualitative comparison results of motion customization on multiple videos.

Tab. 1: Automatic and human evaluations results of motion customization on multiple videos.

		Automatic Evaluations			Human Evaluations			
		Appearance Diversity (↑)	Temporal Consistency (↑)	Pick Score (↑)		Appearance Diversity	Temporal Consistency	Motion Fidelity
Tune-A-Video		28.22	92.45	20.20	v.s. Base Model (ModelScope)	25.00 v.s. 75.00	25.00 v.s. 75.00	40.00 v.s. 60.00
					v.s. Base Model (ZeroScope)	44.00 v.s. 56.00	16.67 v.s. 83.33	53.33 v.s. 46.67
ModelScope	Base Model	28.55	92.54	20.33				
	Coupled Tuned	25.66 (-2.89)	90.66	19.85	v.s. Base Model (ModelScope)	23.08 v.s. 76.92	40.00 v.s. 60.00	52.00 v.s. 48.00
	w/o AD-Loss	28.32 (-0.23)	91.17	20.34	v.s. Base Model (ModelScope)	53.12 v.s. 46.88	49.84 v.s. 50.16	62.45 v.s. 37.55
	ours	**28.66 (+0.11)**	92.36	20.59	v.s. Base Model (ModelScope)	**54.84 v.s. 45.16**	56.00 v.s. 44.00	**75.00 v.s. 25.00**
ZeroScope	Base Model	28.40	92.94	20.76				
	Coupled Tuned	25.52 (-2.88)	90.67	19.99	v.s. Base Model (ZeroScope)	37.81 v.s. 62.19	41.67 v.s. 58.33	54.55 v.s. 45.45
	w/o AD-Loss	28.61 (+0.21)	91.37	20.56	v.s. Base Model (ZeroScope)	50.10 v.s. 49.90	48.00 v.s. 52.00	58.33 v.s. 41.67
	ours	**28.94 (+0.54)**	92.67	20.80	v.s. Base Model (ZeroScope)	**52.94 v.s. 47.06**	55.00 v.s. 45.00	**76.47 v.s. 23.53**

The motion concepts can be ones preferred by users or ones that lie in the long-tailed distribution that can not be synthesized well by pre-trained models. Since appearance and motion are decoupled by our method, the spatial LoRAs can also be used to influence the appearance of generated videos, as shown in Fig. 2. Users can flexibly adjust the influence strength of learned appearance and motion on the generation according to their preferences by simply setting the strength of LoRAs as $W = W_0 + \gamma \Delta W$, where γ is called the LoRA scale, and ΔW is the learned weights. For instance, a small scale of temporal LoRAs leads to the generated results weakly expressing the learned motion concepts, while a significantly large scale leads to discordant generation. More experimental results showing the effect of LoRA scales are provided in the supplementary materials. To ensure comparative fairness, we set the scale of temporal LoRAs to 1 in all experiments in Sec. 4.

4 Experiments

4.1 Motion customization on multiple videos

Dataset. We conduct experiments on the adapted UCF Sports Action data set [53], which includes 95 videos of 12 different human motions, like playing golf, lifting weights, etc. For each type of motion, we label one original text prompt describing the motion, such as "a person is playing golf, side view". For these motions, we set 72 different text prompts in total as input to generate videos using comparison methods, such as "a monkey is playing golf, side view".

Comparison Methods. We compare the proposed method with three baselines and the video generation method Tune-A-Video [70] that can be adapted to this task. Tune-A-Video was initially proposed for training temporal modules on a single video to learn its motion information, while here we adapt it to train on multiple videos. The baseline methods are compared with the proposed method on two different foundational text-to-video diffusion models, i.e. the ModelScope [60] and the ZeroScope [55]. We employ three baseline methods: the first is directly applying the vanilla foundation models, the second is tuning the foundation models with LoRAs in a coupled manner, and the third is the proposed dual-path method excluding the appearance-debiased temporal loss.

Qualitative Results As shown in Fig. 6, taking a set of videos with motions of playing golf as training data, the Tune-A-Video fails to generate diverse appearances with the learned motions, like a monkey playing golf. To compare the baseline methods and proposed method fairly, we feed the same initial Gaussian noise to these methods to generate videos. The pre-trained foundation model, ZeroScope, correctly generates the appearance but lacks the realistic motion that swings a golf club, as those desired motions in the reference videos. The coupled tuned model could generate the desired motion but the learned motion is coupled with too much appearance information causing the generated subject in the video to be more like a human rather than a monkey. The last two rows show that the proposed dual-path LoRAs can avoid hurting the appearance generation and the proposed appearance-debiased temporal loss enhances the learning of desired motion better. We could draw a similar conclusion from the second example showing the motion of riding a panda.

Quantitative Results. We evaluate the methods with automatic evaluations and human evaluations, and the results are shown in Table. 1.

Automatic Metrics. Following the LOVEU-TGVE competition [69], the appearance diversity is computing the average CLIP score [20] between the diverse text prompts and all frames of the generated videos, the temporal consistency is the average CLIP score between frames, and the Pick Score is the average PickScore [31] between all frames of output videos.

In Table 1, the automatic evaluation of temporal consistency is calculated as the CLIP score between frames. This may result in videos that are nearly static and without desired motion receiving high scores. Thus, we further apply human evaluation to evaluate generated videos more comprehensively and accurately.

Human Preference. On the Amazon MTurk [1], each generated video is evaluated by 5 human raters in terms of appearance diversity, temporal consistency, and motion fidelity, which evaluate whether the generated motion is similar to the references. To simplify the comparison for raters, they are asked to compare the results pairwise and select their preferred one, where the videos are shuffled and their source methods are anonymous. In Table. 1, the pairwise numbers "p_1 v.s. p_2" means $p_1\%$ results of the first method are preferred while $p_2\%$ results of the second method are preferred.

The evaluation results show that coupled tuning will destroy the appearance diversity of pre-trained models, while our method will preserve it and achieve the highest motion fidelity.

4.2 Motion customization on a single video

Dataset. We conduct the comparison experiments on the open-sourced benchmark released by the LOVEU-TGVE competition at CVPR 2023 [69]. The dataset comprises 76 videos, each originally associated with 4 editing text prompts. Additionally, we introduced 3 more prompts with significant changes.

Comparison Methods. We compare the proposed method with SOTA controllable generation methods, the VideoCrafter [17], VideoComposer [63], and Control-A-Video [6], and the tuning-based method Tune-A-Video [70]. To ensure a fair comparison, we use the depth control mode of controllable generation methods, which is available in all of them.

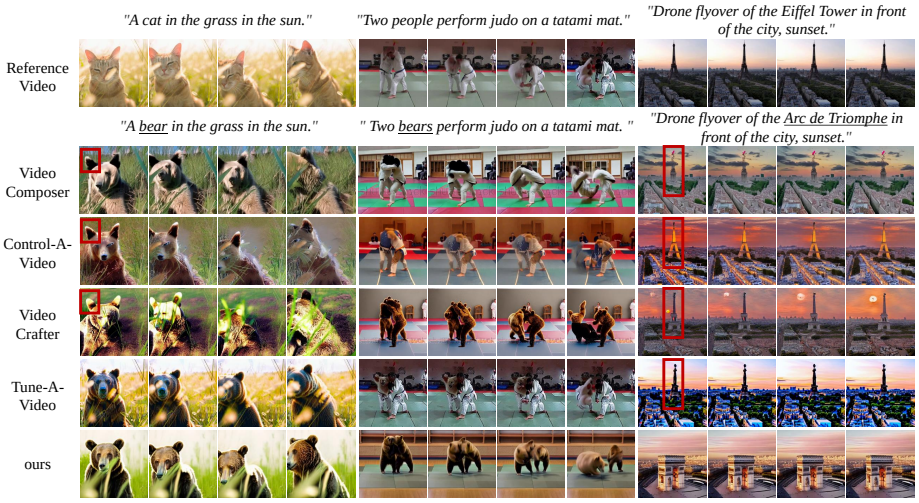

Fig. 7: Qualitative comparison results of motion customization on a single video.

[1] https://requester.mturk.com/

Tab. 2: Automatic and human evaluations results of motion customization on a single video.

Automatic Evaluations				
	Text Alignment (↑)	Appearance Diversity (↑)	Temporal Consistency (↑)	Pick Score (↑)
VideoComposer	27.66	27.03	92.22	20.26
Control-a-Video	26.54	25.35	92.63	19.75
VideoCrafter	**28.03**	27.69	92.26	20.12
Tune-a-Video	25.64	25.95	92.42	20.09
ours	27.82	**28.48**	**93.00**	**20.74**

Human Evaluations				
	Text Alignment	Appearance Diversity	Temporal Consistency	Motion Fidelity
ours v.s. VideoComposer	54.55 v.s. 45.45	72.83 v.s. 27.17	61.57 v.s. 38.43	61.24 v.s. 38.76
ours v.s. Control-A-Video	68.00 v.s. 32.00	78.43 v.s. 21.57	71.28 v.s. 29.72	56.47 v.s. 43.53
ours v.s. VideoCrafter	52.72 v.s. 47.28	71.11 v.s. 28.89	60.22 v.s. 39.78	60.00 v.s. 40.00
ours v.s. Tune-A-Video	67.86 v.s. 32.14	69.14 v.s. 30.86	71.67 v.s. 28.33	56.52 v.s. 43.48

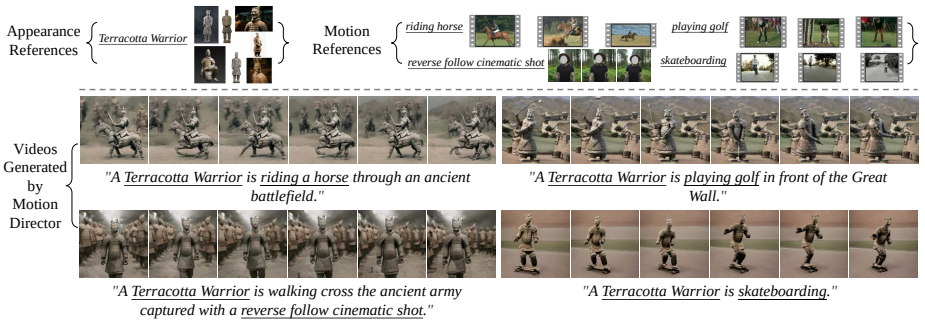

Fig. 8: Videos generated by MotionDirector. Their appearances are customized according to the reference images shown on the top-left side. Their motions are customized based on the reference videos shown on the top-right side.

Qualitative and Quantitative Results. As shown in Fig. 7, comparison methods fail to generalize the desired motions to diverse appearances, like the ears of bears and the Arc de Triomphe. Similar to the depth, using the edge and optical flow as control signals also introduces interference from the appearance into the generated videos. In Table. 2, we refer to the alignment between the generated videos and the original 4 editing text prompts as text alignment, and the alignment with the 3 new text prompts with significant changes as appearance diversity. The results show that our method outperforms other methods by a large margin when generalizing the motions to diverse appearances, and achieves competitive motion fidelity.

4.3 Customizing Both Appearance and Motion

Since MotionDirector learns the appearance and motion in a decoupled manner, it can also reorganize and combine the appearances and motions from different sources. For example, as shown in Fig. 2, MotionDirector can combine the appearance from one video with the motion from another video to generate a new video that exhibits similar appearance and motion characteristics to each of the input videos, respectively. Here we provide more results in Fig. 8. The base model is injected with the spatial LoRAs trained on the given images of "Terracotta Warrior" and injected with different temporal LoRAs trained on different sets of videos of different motion concepts, such as "riding horse".

4.4 Efficiency Performance

The lightweight LoRAs enable our method to tune the foundation models efficiently. Taking the foundation model ZeroScope for example, it has over 1.8 billion pre-trained parameters. Each set of trainable spatial and temporal LoRAs only adds 9 million and 12 million parameters, respectively. Requiring 14 GB VRAM, MotionDirector takes 20 minutes to converge on multiple reference videos, and 8 minutes for a single reference video, competitive to the 10 minutes required by Tune-A-Video [70]. After training, MotionDirector can generate diverse videos with learned motions. In the inference stage, MotionDirector does not introduce significant increases in computational costs or inference time.

5 Limitations and Future Works

Despite the MotionDiector can learn the motions of one or two subjects in the reference videos, it is still hard to learn complex motions of multiple subjects, such as a group of boys playing soccer. Previous appearance customization methods suffer similar problems when generating multiple customized subjects [14]. A possible solution is to further decouple the motions of different subjects in the latent space and learn them separately.

6 Conclusion

We introduce and formulate the task of Motion Customization, which is adapting the pre-trained foundation text-to-video diffusion models to generate videos with desired motions. The challenge of this task is generalizing the customized motions to various appearances. To overcome this challenge, we propose the MotionDirector with a dual-path architecture and a novel appearance-debiased temporal training objective to decouple the learning of appearance and motion. Experimental results show that MotionDirector can learn motion concepts from a limited number of reference videos, and generalize them to diverse appearances. The automatic and human evaluations on two benchmarks demonstrate the MontionDirector outperforms other methods in terms of appearance diversity and motion fidelity.

Acknowledgements

This research is supported by National Research Foundation, Singapore and A*STAR, under its RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) grant call (Grant No. I2001E0059) – SIA-NUS Digital Aviation Corp Lab. Mike Shou is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2023).

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738 (2021)
2. Balaji, Y., Min, M.R., Bai, B., Chellappa, R., Graf, H.P.: Conditional gan with discriminative filter generation for text-to-video synthesis. In: IJCAI. vol. 1, p. 2 (2019)
3. 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: CVPR (2023)
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. Chen, T.S., Lin, C.H., Tseng, H.Y., Lin, T.Y., Yang, M.H.: Motion-conditioned diffusion model for controllable video synthesis. arXiv:2304.14404 (2023)
6. Chen, W., Wu, J., Xie, P., Wu, H., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video generation with diffusion models. arXiv preprint arXiv:2305.13840 (2023)
7. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
10. Duan, Z., You, L., Wang, C., Chen, C., Wu, Z., Qian, W., Huang, J., Chao, F., Ji, R.: Diffsynth: Latent in-iteration deflickering for realistic video synthesis. arXiv:2308.03463 (2023)
11. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. arXiv preprint arXiv:2302.03011 (2023)
12. 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. arXiv preprint arXiv:2204.03638 (2022)
13. 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. arXiv:2305.10474 (2023)
14. Gu, Y., Wang, X., Wu, J.Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al.: Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. arXiv preprint arXiv:2305.18292 (2023)
15. 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)
16. He, Y., Xia, M., Chen, H., Cun, X., Gong, Y., Xing, J., Zhang, Y., Wang, X., Weng, C., Shan, Y., et al.: Animate-a-story: Storytelling with retrieval-augmented video generation. arXiv:2307.06940 (2023)

17. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation (2022)
18. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv:2211.13221 (2022)
19. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Videocrafter: A toolkit for text-to-video generation and editing. `https://github.com/AILab-CVC/VideoCrafter` (2023)
20. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
21. 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)
22. 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:2210.02303 (2022)
23. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: NeurIPS (2022)
24. Hong, S., Seo, J., Hong, S., Shin, H., Kim, S.: Large language models are frame-level directors for zero-shot text-to-video generation. arXiv:2305.14330 (2023)
25. 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)
26. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
27. Huang, H., Feng, Y., Shi, C., Xu, L., Yu, J., Yang, S.: Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. In: NeurIPS (2023)
28. Jeong, H., Park, G.Y., Ye, J.C.: Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. arXiv preprint arXiv:2312.00845 (2023)
29. Karras, J., Holynski, A., Wang, T.C., Kemelmacher-Shlizerman, I.: Dreampose: Fashion image-to-video synthesis via stable diffusion. arXiv:2304.06025 (2023)
30. 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. In: ICCV (2023)
31. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. arXiv preprint arXiv:2305.01569 (2023)
32. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
33. Le Moing, G., Ponce, J., Schmid, C.: Ccvs: context-aware controllable video synthesis. Advances in Neural Information Processing Systems **34**, 14042–14055 (2021)
34. 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)
35. Lian, L., Shi, B., Yala, A., Darrell, T., Li, B.: Llm-grounded video diffusion models. arXiv:2309.17444 (2023)
36. 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 gen-

eration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
37. Ma, Y., He, Y., Cun, X., Wang, X., Shan, Y., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. arXiv preprint arXiv:2304.01186 (2023)
38. Ma, Y., He, Y., Cun, X., Wang, X., Shan, Y., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. arXiv:2304.01186 (2023)
39. Materzynska, J., Sivic, J., Shechtman, E., Torralba, A., Zhang, R., Russell, B.: Customizing motion in text-to-video diffusion models. arXiv preprint arXiv:2312.04966 (2023)
40. Mei, K., Patel, V.: Vidm: Video implicit diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 9117–9125 (2023)
41. Ni, H., Shi, C., Li, K., Huang, S.X., Min, M.R.: Conditional image-to-video generation with latent flow diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18444–18455 (2023)
42. Qin, B., Ye, W., Yu, Q., Tang, S., Zhuang, Y.: Dancing avatar: Pose and text-guided human motion videos synthesis with image diffusion model. arXiv:2308.07749 (2023)
43. Ren, Y., Zhou, Y., Yang, J., Shi, J., Liu, D., Liu, F., Kwon, M., Shrivastava, A.: Customize-a-video: One-shot motion customization of text-to-video diffusion models. arXiv preprint arXiv:2402.14780 (2024)
44. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
45. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
46. Ryu, S.: Low-rank adaptation for fast text-to-image diffusion fine-tuning. `https://github.com/cloneofsimo/lora` (2023)
47. 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)
48. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
49. Shen, X., Li, X., Elhoseiny, M.: Mostgan-v: Video generation with temporal motion styles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5652–5661 (2023)
50. 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)
51. Skorokhodov, I., Tulyakov, S., Elhoseiny, M.: Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. arXiv preprint arXiv:2112.14683 (2021)
52. Smith, J.S., Hsu, Y.C., Zhang, L., Hua, T., Kira, Z., Shen, Y., Jin, H.: Continual diffusion: Continual customization of text-to-image diffusion with c-lora. arXiv preprint arXiv:2304.06027 (2023)
53. Soomro, K., Zamir, A.R.: Action recognition in realistic sports videos. In: Computer vision in sports, pp. 181–208. Springer (2015)

54. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: International conference on machine learning. pp. 843–852. PMLR (2015)
55. Sterling, S.: Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w (2023)
56. Tian, Y., Ren, J., Chai, M., Olszewski, K., Peng, X., Metaxas, D.N., Tulyakov, S.: A good image generator is what you need for high-resolution video synthesis. In: International Conference on Learning Representations (2020)
57. 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)
58. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. Advances in Neural Information Processing Systems **35**, 23371–23385 (2022)
59. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. Advances in neural information processing systems **29** (2016)
60. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
61. Wang, T., Li, L., Lin, K., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. arXiv:2307.00040 (2023)
62. Wang, W., Yang, H., Tuo, Z., He, H., Zhu, J., Fu, J., Liu, J.: Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. arXiv:2305.10874 (2023)
63. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. arXiv preprint arXiv:2306.02018 (2023)
64. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., der Yang, P., Guo, Y., Wu, T., Si, C., Jiang, Y., Chen, C., Loy, C.C., Dai, B., Lin, D., Qiao, Y., Liu, Z.: Lavie: High-quality video generation with cascaded latent diffusion models (2023), https://api.semanticscholar.org/CorpusID:262823915
65. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. arXiv:2309.15103 (2023)
66. Wang, Z., Yuan, Z., Wang, X., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. arXiv preprint arXiv:2312.03641 (2023)
67. Wei, Y., Zhang, S., Qing, Z., Yuan, H., Liu, Z., Liu, Y., Zhang, Y., Zhou, J., Shan, H.: Dreamvideo: Composing your dream videos with customized subject and motion. arXiv preprint arXiv:2312.04433 (2023)
68. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848 (2023)
69. Wu, J.Z., Gao, D., Bai, J., Shou, M., Li, X., Dong, Z., Singh, A., Keutzer, K., Iandola, F.: The text-guided video editing benchmark at loveu 2023. https://sites.google.com/view/loveucvpr23/track4 (2023)
70. Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, 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. arXiv preprint arXiv:2212.11565 (2022)

71. Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, 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: ICCV (2023)
72. Wu, R., Chen, L., Yang, T., Guo, C., Li, C., Zhang, X.: Lamp: Learn a motion pattern for few-shot-based video generation. arXiv preprint arXiv:2310.10769 (2023)
73. Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., et al.: Make-your-video: Customized video generation using textual and structural guidance. arXiv:2306.00943 (2023)
74. Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., Wu, Z., Jiang, Y.G.: A survey on video diffusion models. arXiv preprint arXiv:2310.10647 (2023)
75. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
76. Yang, S., Hou, L., Huang, H., Ma, C., Wan, P., Zhang, D., Chen, X., Liao, J.: Direct-a-video: Customized video generation with user-directed camera movement and object motion. arXiv preprint arXiv:2402.03162 (2024)
77. Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv:2308.08089 (2023)
78. Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., et al.: Nuwa-xl: Diffusion over diffusion for extremely long video generation. arXiv:2303.12346 (2023)
79. 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)
80. Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J.W., Shin, J.: Generating videos with dynamics-aware implicit generative adversarial networks. In: International Conference on Learning Representations (2021)
81. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. arXiv:2309.15818 (2023)
82. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation (2023)
83. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)
84. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. arXiv preprint arXiv:2310.08465 (2023)
85. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. arXiv preprint arXiv:2305.16322 (2023)
86. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv:2211.11018 (2022)