# MotionCrafter: Plug-and-play Motion Guidance for Diffusion Models

Yuxin Zhang, Weiming Dong, *Member, IEEE*, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Pengfei Wan, Tong-Yee Lee, *Senior Member, IEEE* Changsheng Xu, *Fellow, IEEE*

**Abstract**—The essence of a video lies in the dynamic motions. While text-to-video generative diffusion models have made significant strides in creating diverse content, effectively controlling specific motions through text prompts remains a challenge. By utilizing user-specified reference videos, the more precise guidance for character actions, object movements, and camera movements can be achieved. This gives rise to the task of motion customization, where the primary challenge lies in effectively decoupling the appearance and motion within a video clip. To address this challenge, we introduce *MotionCrafter*, a novel one-shot instance-guided motion customization method that is suitable for both pre-trained text-to-video and text-to-image diffusion models. *MotionCrafter* employs a parallel spatial-temporal architecture that integrates the reference motion into the temporal component of the base model, while independently adjusting the spatial module for character or style control. To enhance the disentanglement of motion and appearance, we propose an innovative dual-branch motion disentanglement approach, which includes a motion disentanglement loss and an appearance prior enhancement strategy. To facilitate more efficient learning of motions, we further propose a novel timestep-layered tuning strategy that directs the diffusion model to focus on motion-level information. Through comprehensive quantitative and qualitative experiments, along with user preference tests, we demonstrate that *MotionCrafter* can successfully integrate dynamic motions while maintaining the coherence and quality of the base model, providing a wide range of appearance generation capabilities. *MotionCrafter* can be applied to various personalized backbones in the community to generate videos with a variety of artistic styles.

**Index Terms**—Text-to-video generation, diffusion models, motion generation

## 1 Introduction

Dynamic scenes that resonate with our emotions are not solely memorable due to their captivating visual appeal, but also through the compelling performances of actors, engaging storylines, and meticulous cinematography. Unlike images, videos encapsulate both spatial and temporal information, resulting in unique dynamic motions. While current text-to-video techniques can generate videos based on user-input text [2]–[4], specific information about actions, object movements, and camera movements in the generated videos often cannot be accurately described by text. Consequently, a significant challenge persists: how can we effectively leverage existing pre-trained models and allow users to precisely customize various temporal aspects in videos.

Several current methods, such as textual inversion (TI) [5], adopt a personal concept representation technique that maps image references to a text-conditioned space. This integration with natural language aids in instance-guided concept replication and manipulation. Other approaches, such as DreamBooth [1] and Custom Diffusion [6], focus on fine-tuning the core model's internal parameters. They leverage a set of images provided by the user to augment the model's ability to express specific concepts more effectively. These aforementioned approaches have demonstrated effective performance in tasks involving instance-guided visual customization. In addition, various video generation methods [7]–[10] have leveraged them to control artistic styles or specific characters in videos. However, these image-centric concept representation techniques mainly emphasize appearance, overlooking the essential temporal attributes unique to videos. Consequently, the specific challenge of customizing motion in videos remains unaddressed by these existing approaches.

To enhance controllability and expressiveness, inspired by the customization of text-to-image generation, the concept of motion customization is naturally introduced. Motion customization aims to offer users more precise control over actions, object movements, and camera movements, by allowing them to specify target motions through video inputs. The primary challenge is to proficiently learn and represent these visual content, requiring the disentanglement and manipulation of temporal elements and network components within existing text-to-video generation framework.

In this work, we introduce *MotionCrafter*, a novel one-shot instance-guided method for dynamic motion customization of both pre-trained T2V models and various T2I personalization models. To learn customization motion

- *Yuxin Zhang, Weiming Dong, Nisha Huang, and Changsheng Xu are with MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100040, China. E-mail: {zhangyuxin2020, weiming.dong, huangnisha2021,}@ia.ac.cn and csxu@nlpr.ia.ac.cn.*
- *Fan Tang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100040, China. E-mail: tfan.108@gmail.com.*
- *Haibin Huang, Chongyang Ma, and Pengfei Wan are with Visual Generation and Interaction Center, Kuaishou Technology, Beijing 100085, China. E-mail: jackiehuanghaibin@gmail.com, chongyangm@gmail.com and wanpengfei@kuaishou.com.*
- *Tong-Yee Lee is with National Cheng Kung University, Tainan 701, Taiwan. E-mail: tonylee@mail.ncku.edu.tw.*

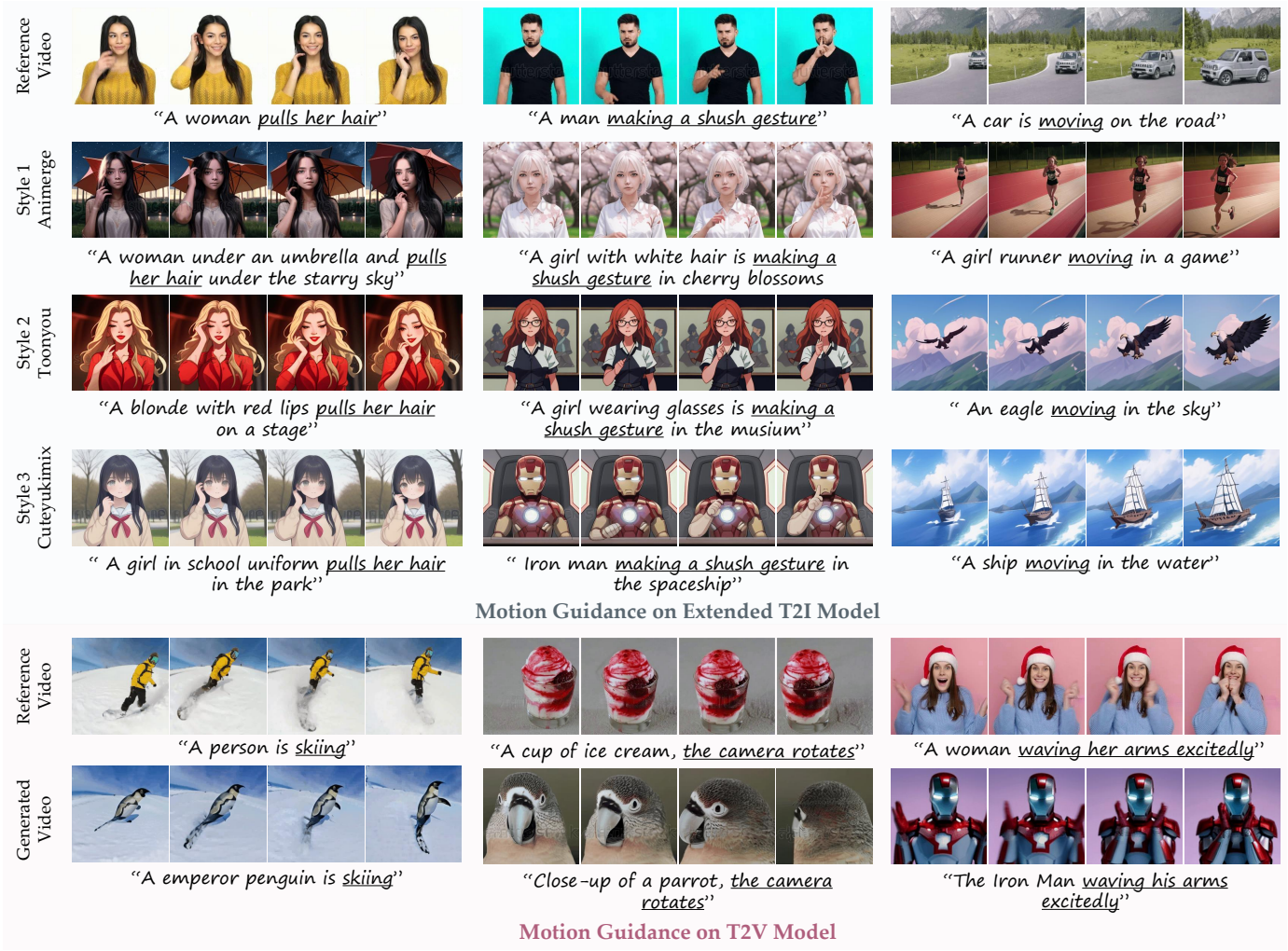



Fig. 1: Motion customization results using *MotionCrafter*. Our method can effectively disentangle the motion features from the input video and generate new videos with the same motion but different objects from text prompts or with personalized models. With various personalized backbones in the community, *MotionCrafter* can produce videos with a variety of artistic styles. Animerge, Cuteyukimix, and ToonYou represent 2D stylized text-to-image generation models fine-tuned on DreamBooth [1].

concepts from a given video clip, we introduce a parallel spatial-temporal architecture to infuse motion information into the temporal attention module of the model, aiming at customizing video appearance and motion separately. To achieve motion disentanglement, we propose a dual-branch motion disentanglement strategy by introducing the base model as a prior. During the training phase, a frozen U-Net is incorporated with the trainable network and both perform inference on the same text conditions, leading to results of the base model and generated output with injected motions, respectively. Furthermore, we design an appearance prior enhancement scheme where we keep the text prompt for the intended motion fixed, while altering descriptions of various appearances (such as character, scene, etc) during training. This scheme encourages the frozen base U-Net model to generate results with diverse appearances. Then, using our proposed motion disentanglement loss, we regulate the mutual information between the generated and reference videos, as well as between the generated results and those from the base model, removing the appearance information from the reference video to achieve motion-appearance decoupling while preserving the generation capability of the base model. To learn motion more efficiently, we propose a timestep-layered tuning strategy. Leveraging the characteristic of the diffusion model that different timesteps and U-Net layers generate contents of different semantic levels, we enable the model to focus on motion information. Our proposed plug-and-play motion customization method leverages the high-quality personalized models already available in the community, and makes it possible for non-research users, such as artists and hobbyists, to participate in a low-cost way. Our contributions can be summarized as follows:

- We present *MotionCrafter*, a one-shot instance-guided motion customization framework that can plug-and-play with both pre-trained T2V models and various T2I personalization models, enabling low-cost motion control video generation. With various personalized backbones in the community, *MotionCrafter* can

produce videos with a variety of artistic styles.
- We propose a dual-branch motion disentanglement approach, which facilitates the separation of motion and appearance by producing diverse appearance priors from the base model.
- We propose a novel timestep-layered tuning strategy to force the diffusion model focusing on motion-level information, achieving more efficient and accurate motion learning.
- Experiments on a wide range of customized motions, including object movements, human actions, and camera movements, demonstrate the proposed *MotionCrafter* achieves state-of-the-art performance and also generate appealing video results.

## 2 Related Work

### 2.1 Text-to-video synthesis and editing

Recent studies [2], [3], [11] have employed diffusion models to generate lifelike videos, harnessing text as a powerful guiding instruction. VideoFusion [4] employs a decomposition-diffusion process to enhance control over content and motion in video generation. Furthermore, Imagen Video [12] explores the effectiveness of v-prediction parameterization on sample quality and the progressive distillation-guided diffusion model in video generation. Similarly, VideoFactory [13] introduces a novel swapped spatial-temporal cross-attention mechanism that reinforces spatial and temporal interactions. ModelScopeT2V [14] employs spatial-temporal blocks and a multi-frame training strategy to effectively model temporal dependencies, ensuring smooth motion between frames and improving performance and generalization by learning motion patterns from image-text and video-text datasets.

The above methods care about the consistency between the text and the generated video, while accurately controlling the specific dynamics of the generated video is still challenging. Several methods aim at controlling videos. ControlVideo [15] is a training-free framework for text-to-video generation using fully cross-frame interaction in self-attention modules and the generated motions are controlled by edge or depth maps. Control-A-Video [16] incorporates a spatial-temporal self-attention mechanism into the text-to-image diffusion model, enabling video generation based on control signal sequences. Rerender-A-Video [17] incorporates hierarchical cross-frame constraints and employs time-aware patch matching and frame blending, maintaining shape, texture, and color consistency in the translated video. Tune-A-Video [18] introduces a one-shot video tuning method to achieve video editing. The above methods focus on controlling the structure of the video rather than the dynamic information. VideoComposer [19] is a textual-spatial-temporal controllable video generation method. They introduce 2D motion vectors that capture pixel-wise movements between adjacent frames, as an explicit control signal to guide temporal dynamics. However, motion vectors are difficult to deal with complex movements and large shape changes. Make-A-Video [20], Align-your-Latents [21], and AnimateDiff [7] introduce different temporal modules to the latent diffusion model and transform the image generator into a video generator. They obtain inter-frame consistency by training the temporal module using extensive video data. They further propose AnimateDiff v2 which includes a motion LoRA [22] to learn camera movements. We inject specific motions, includes camera movements, human actions and object movements, into the temporal module, thereby achieving more precise control.

### 2.2 Customization of generative models

The customization of text-to-image and text-to-video generation models involves the learning of personalized concept with pre-trained models. Gal et al. [5] propose the task of textual inversion that aims to find a pseudo-word that describes the visual concept of a specific object in a set of user-provided images. Avrahami et al. [23] introduce a method to extract distinct text tokens for each concept from images containing multiple concepts. Besides objects, several methods aim to learn different concepts from given images. Zhang et al. [24] propose InST, an attention-based inversion style transfer method. Huang et al. [25] propose ReVersion for relation inversion, with the aim of learning specific relations from images. Wen et al. [26] introduce the concept of hard prompts, which invert a given concept into readable natural languages. Voynov et al. [27] present an extended textual conditioning space consisting of several textual embeddings derived from per-layer prompts, each corresponding to a layer of diffusion model's denoising U-Net. Zhang et al. [28] reveal that diffusion models generate images by prioritizing low to high frequency information and represent images as a compilation of inverted textual token embeddings generated from per-stage prompts. Instead of inverting a concept into textual tokens, DreamBooth [1] generates a specific subject by finetuning the diffusion models with a unique identifier. Kumari et al. [6] propose Custom Diffusion, which optimizes a few parameters in the conditioning mechanism and can be jointly trained for multiple concepts or combine several fine-tuned models. Chen et al. [29] propose AnyDoor, a diffusion-based image generator that can teleport target objects to new scenes at specified locations. The above methods focus on concept learning in image generation, while we aim to control motions in video generation.

For text-to-video generation, Gong et al. [9] introduce TaleCrafter, an interactive story creation system that can handle multiple characters with layout and structural editing capabilities. He et al. [8] propose a retrieval-based depth-guided method that leverages existing video clips to create a coherent storytelling video by customizing the appearances of characters. These methods focus on modeling the appearance of visual content, but have difficulty controlling motions. In contrast, our approach emphasizes the specific inter-frame dynamics of the video. Zhao et al. [30] introduce MotionDirector, which learns motions through customized diffusion models. Wei et al. [31] introduce DreamVideo which can personalize the character and the motion by tuning the diffusion models. Jeong et al. [32] introduce VMC, a motion distillation objective that employs residual vectors between consecutive frames as a motion reference. PIA [33] introduces an image-to-video motion generation approach that modulates motion using textual prompts rather than reference videos. HiGen [34] proposes a two-stage video generation framework, encompassing a text-to-image appearance generation stage and an image-to-video

motion generation stage. Direct-a-Video [35] integrates an additional bounding box to control object movement. While this approach effectively regulates object positioning, it remains limited in achieving precise control over specific actions. Our method excels in one-shot motion learning, necessitating a meticulous disentanglement of motion and appearance elements. Additionally, we demonstrate that our method is compatible with T2I models, thereby enhancing its versatility.

## 3 Our Method

The overall pipeline of *MotionCrafter* is illustrated in Fig. 2. To decompose the appearance and motion of the generated videos, we propose a parallel spatial-temporal architecture (Section 3.1). It leverages two separate paths to learn the appearance and motion information from videos, corresponding to the spatial and temporal modules in the backbone of a text-to-video generation model. To achieve better disentanglement, we further design a dual-branch motion disentanglement based on an information bottleneck (Section 3.2). We incorporate a frozen branch of the base model to serve as an appearance prior. To better capture the motion, which appears in structure or layout changes, we propose a novel timed-layered tuning (Section 3.2.3). During training, the framework takes a reference video and enhances textual conditioning as inputs and fine-tunes the trainable branch. During inference, our framework takes user-provided textual conditioning as input and generates results that incorporate the reference video information using only the fine-tuned branch.

### 3.1 Spatial-Temporal Architecture

Our goal is a unified instance-guided motion framework suitable for both pre-trained text-to-video and text-to-image diffusion models. To this end, we employ the widely adopted text-to-video diffusion model [36] and the expanded text-to-image model [7] as the foundational models. The text-to-video diffusion model is structured around a 3D U-Net framework, which operates on the latent space derived from an autoencoder. The U-Net backbone consists of down blocks, mid blocks, and up blocks, each accompanied by spatial and temporal attention and convolutional modules. The spatial attention module performs operations on the 2D spatial dimensions encompassing the width and height of the latent codes. At each timestep $t$, the 3D U-Net takes a latent code $z_t$ with dimensions $batch \times frames \times width \times height \times channels$ as input, and gives the predicted noise $\epsilon_\theta(\cdot)$, where $\theta$ denotes model parameters. Additionally, textual conditions $\tau_\theta(y)$, where $y$ denotes input conditions, are incorporated to provide contextual guidance module captures inter-frame relationships along the frame dimension. The extension of the text-to-image diffusion model is specifically tailored for text-to-video generation. Between the layers of the U-Net, the predicted latents are fed into an extended motion module, which provides cross-frame perception. At each timestep $t$, the text-to-image U-Net backbone takes a latent code $z_t$ with dimensions $batch \times width \times height \times channels$ as input, where $batch$ is used as $frames$. By introducing the motion module within the text-to-image generation framework and training this module on a large video dataset, the expanded text-to-image model is capable of generating coherent videos. Although the motion module provides better inter-frame continuity, it is difficult to generate accurate and high-quality dynamics.

In our approach, we design a spatial-temporal learning framework that leverages the intrinsic properties of the temporal and spatial modules within the two types of foundational models. The spatial attention and convolutional module within the text-to-video models and the text-to-image backbone are named as spatial module. The temporal attention and convolutional module within the text-to-video models and the injected motion module are named as temporal module. For the appearance information, by leveraging techniques from DreamBooth [1] or LoRA [22] with its text-to-image backbone, the extended text-to-image diffusion models can also proficiently produce videos in diverse styles following the training of the motion module.

For the motion information, we fine-tune the temporal modules to update the correlations along the temporal dimension. For T2V and extended T2I models, their temporal modules' structure may be different, and our method should be used on the desired model. The temporal loss is formulated as:

$$\mathcal{L}_{temporal} = \mathbb{E}_{\mathcal{E}\left(x_0^{1:N}\right),y,\epsilon\sim\mathcal{N}(0,I),t}\left[\left\|\epsilon-\epsilon_\theta\left(z_t^{1:N},t,\tau_\theta(y)\right)\right\|_2^2\right], \quad (1)$$

where $N$ denotes the number of frames, and $t$ denotes the sampled diffusion time step. However, during the training and inference stages, the information represented by the spatial and temporal modules is coupled together. Therefore, the challenge of customizing motions based on pre-trained text-to-video generation models lies in decomposing the spatial and temporal information of the generated video.

### 3.2 Dual-Branch Motion Disentanglement

#### 3.2.1 Motion disentanglement loss

To address the aforementioned coupling issue of spatial and temporal information, we introduce a dual-branch framework for motion disentanglement in videos, as illustrated in Figs. 2 and 3. The fitting of the model to the appearance of input videos leads to the inherent loss of its own diversity. By introducing a base model as a prior, the diversity can be better preserved, thereby alleviating the issue of appearance overfitting. During the training process, we incorporate an additional frozen U-Net maintaining the parameters of the base model to provide normalization videos. To separate appearance information from the reference video, we introduce a motion disentanglement loss based on the information bottleneck, which consists of an appearance normalization loss that pushes the generated results to match the normalization videos, and the aforementioned temporal loss encouraging the model to generate results consistent with the reference video. Thus, by controlling the accessibility of information bottlenecks, we can effectively eliminate appearance information from the reference video while avoiding overfitting. This approach ensures the preservation

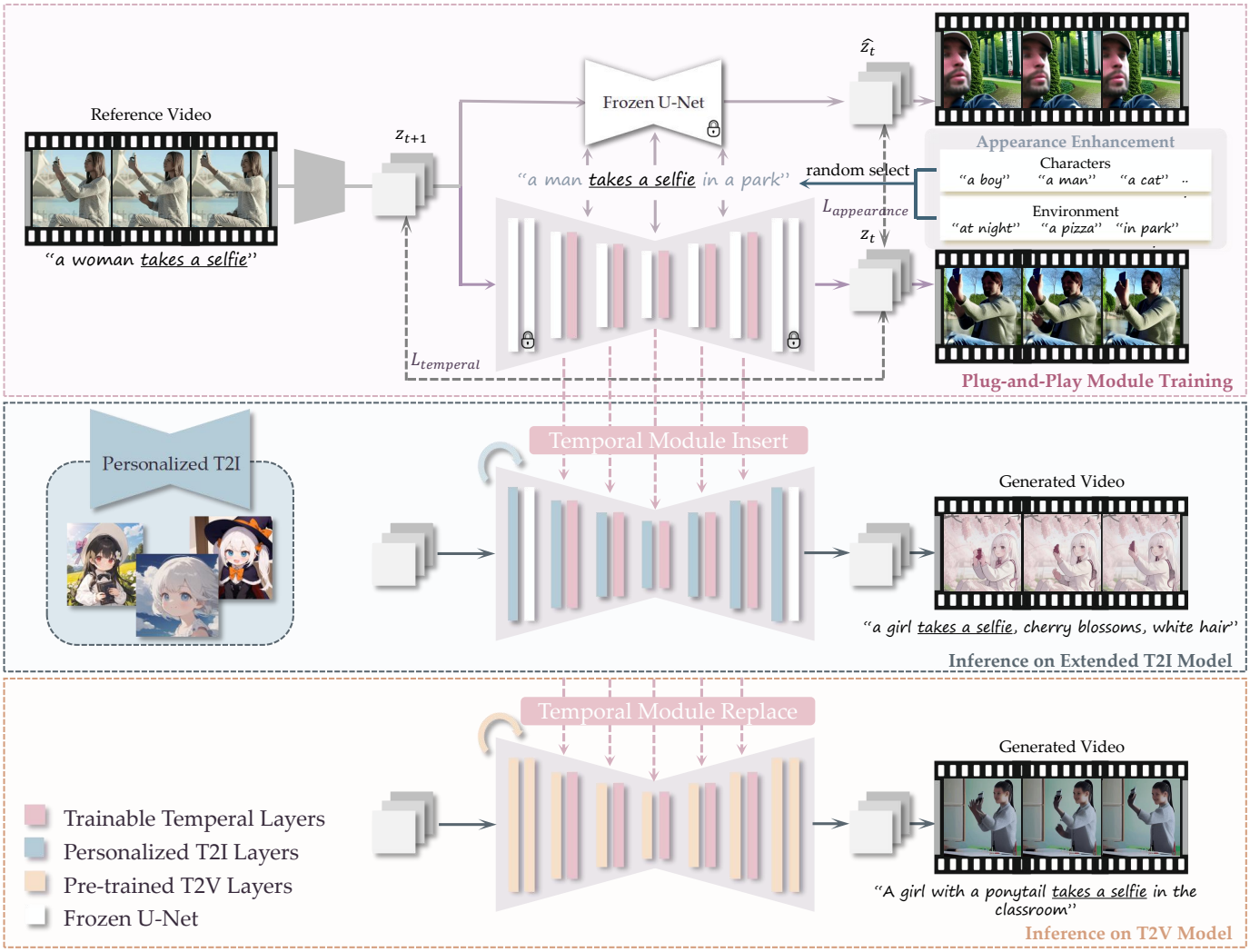



Fig. 2: The overall pipeline of *MotionCrafter*. As shown in the pink box, we use a parallel spatial-temporal architecture to inject appearance and motion into the corresponding layers in a one-shot fine-tuning manner. We introduce a frozen U-Net, which retains the basic model parameters. During training, the temporally-tuned U-Net and the frozen U-Net receive the same text prompt for appearance enhancement, resulting in normalized latents. The appearance regularization loss $\mathcal{L}_{appearance}$ is used to enforce similarity between the two latents, while the temporal loss $\mathcal{L}_{temperal}$ is calculated between the generated and reference latents. As shown in the blue box, *MotionCrafter* can be injected into different personalized text-to-image models and generate consistent motions in various styles. As shown in the orange box, *MotionCrafter* can also be applied to pre-trained text-to-video generation models by replacing temporal modules thus achieving motion control. Each one of the two backbones is used alone with *MotionCrafter*.

of the pre-trained model's appearance diversity, leading to improved controllability of the generated videos.

Specifically, in the autoencoder's latent space, the dual-branch U-Net consists of a frozen U-Net backbone $\hat{\mathcal{U}}$ with the original weights and another U-Net $\mathcal{U}_\theta$ with trainable temporal layers. At each timestep, a shared latent code undergoes separate processing by the two branches, resulting in $\hat{z}_t$ and $z_t$. In this process, $\hat{z}_t$ preserves the diverse appearance generated by the frozen model, while the reference information is injected into $z_t$ via the trainable branch. We propose an appearance normalization loss $\mathcal{L}_{appearance}$, which imposes a constraint on the KL divergence between the distributions of the latent codes $z_t$ and $\hat{z}_t$. By aligning the distributions, the appearance information of the reference video is squeezed out. This process is parameterized as:

$$\mathcal{L}_{appearance} = D_{KL}\left(q_\theta(z_t \mid x_t) \| p(\hat{z}_t)\right), \tag{2}$$

where $q$ and $p$ denote different distributions. The appearance normalization loss described above is combined with a temporal loss $\mathcal{L}_{temperal}$, thereby facilitating the extraction of motion. The full objective is named as our motion disentanglement loss $\mathcal{L}_{motion}$ and is formulated as:

$$\mathcal{L}_{motion} = \mathcal{L}_{temperal} + \beta \mathcal{L}_{appearance}, \tag{3}$$

where the hyper-parameter $\beta$ is included to control the accessibility of information bottlenecks. We set $\beta = 5$ in all the experiments.

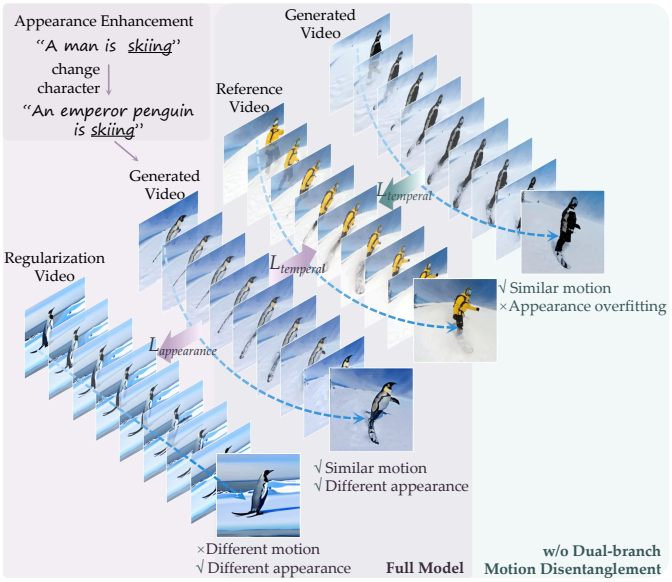



Fig. 3: The workflow of Dual-branch Motion Disentanglement enables the model to strike a balance between the reference video and the normalization video. Thus, *MotionCrafter* can maintain the base model's appearance while have a consistent motion as the reference video. Without the motion disentanglement, as illustrated on the right, one-shot customization could lead to overfitting of the appearance to the reference videos.

#### 3.2.2 Appearance prior enhancement scheme

To better preserve the diversity of appearance in the generated backbone, we propose an appearance prior enhancement scheme. This scheme is designed to encourage the original U-Net model to generate more diverse content while simultaneously preserving the intended motion. Specifically, we create templates that capture a variety of appearances, including descriptions of diverse objects and scenes in natural language (e.g., "a woman with a hat is {} in a park"). By incorporating the appearance details of these templates and combining them with the target motion (i.e., replacing with a description of the motion), the original U-Net $\hat{\mathcal{U}}$, which remains frozen, is able to generate results with a more diverse appearance while maintaining motion information. This scheme helps the network differentiate between appearance and motion information, leading to a more comprehensive disentanglement.

#### 3.2.3 Timestep-Layered Tuning

**Timestep-related Tuning.** Our goal is to achieve a plug-and-play motion module that learns high-quality motions while minimizing the impact on the expressiveness of the generative backbones. The diffusion model has been revealed to generate images in the order of structure first and then details [28], [39] in the dimension of denoised timesteps. The motion information within a video can be classified as layout and structured information, which is generated in the early stage of the diffusion model. Artistic styles, e.g., brushstrokes and textures, belong to fine-grained information and are often generated in the later stages. This inspires a novel timestep-related tuning strategy in our paper. During each iteration of the prior training, a single step is sampled from the whole steps of the diffusion process to uniformly optimize the entire model. Timestep-related tuning ensures that the model remains focused on target attributes by influencing the fluency of each sampled timestep. Specifically, this approach represents a novel sampling strategy for timesteps during the fine-tuning process. It is a versatile strategy applicable to various tasks aimed at learning specific attributes and is adopted in several previous methods [25] in image generation. For motion customization, timestep-related tuning is designed to focus on the early stages of generation to better learn motion dynamics. We set the denoising probability of steps 0-500 to 0.8 and the sampling probability of steps 500-1000 to 0.2 to focus the model on structure-related generation. Thus, time-tuning increases the accuracy of motion learning, and reduces the impact of appearance overfitting on the generated styles.

**U-Net layered tuning.** The motion module should efficiently and effectively learn motion information while reducing interference from the original video's appearance. Voynov et al. [27] demonstrate that the U-Net structure of the diffusion model tends to generate different types of content at different layers, with deep layers capturing the shape and structural information and shallow layers expressing color and texture information. This inspires us to utilize a U-Net layered tuning approach, which is used in some image generation methods [29], [40], [41]. The motion within a video can be seen as structured information, generated in the deep layers of the U-Net, while appearance is generated in the shallow layers. Only the parameters of the the desired information in the target layers are optimized, while the other parameters are frozen. Specifically, as shown in Fig. 2, we optimize the deep layers of the U-Net using the motion disentanglement loss, allowing the model to better focus on the motion information while keeping the shallow layers frozen to reduce appearance overfitting.

## 4 Experiments

In this section, we demonstrate that *MotionCrafter* is capable of replicating motions from the reference video while maintaining content coherence. Additionally, it offers greater editability compared to state-of-the-art text-to-video customization baselines.

### 4.1 Experimental Setup

**Methods for comparison.** We compare our approach with state-of-the-art text-to-video generation methods ZeroScope [36] and AnimateDiff [7], several video-to-video editing methods including Control-A-Video [16], VideoComposer [19], ControlVideo [15], as well as the fine-tuning-based video customization method Tune-A-Video [18],MotionDirector [30],VMC [32] and DreamVideo [31].

**Evaluation dataset.** To ensure a fair comparison, we utilize widely used video segments from previous papers, along with clips from the WebVid-10M dataset [42], UCF Sports dataset [37] and DAVIS dataset [38]. For WebVid-10M and DAVIS datasets, we use 20 motions for qualitative and quantitative evaluations. For UCF Sports dataset, we use 8

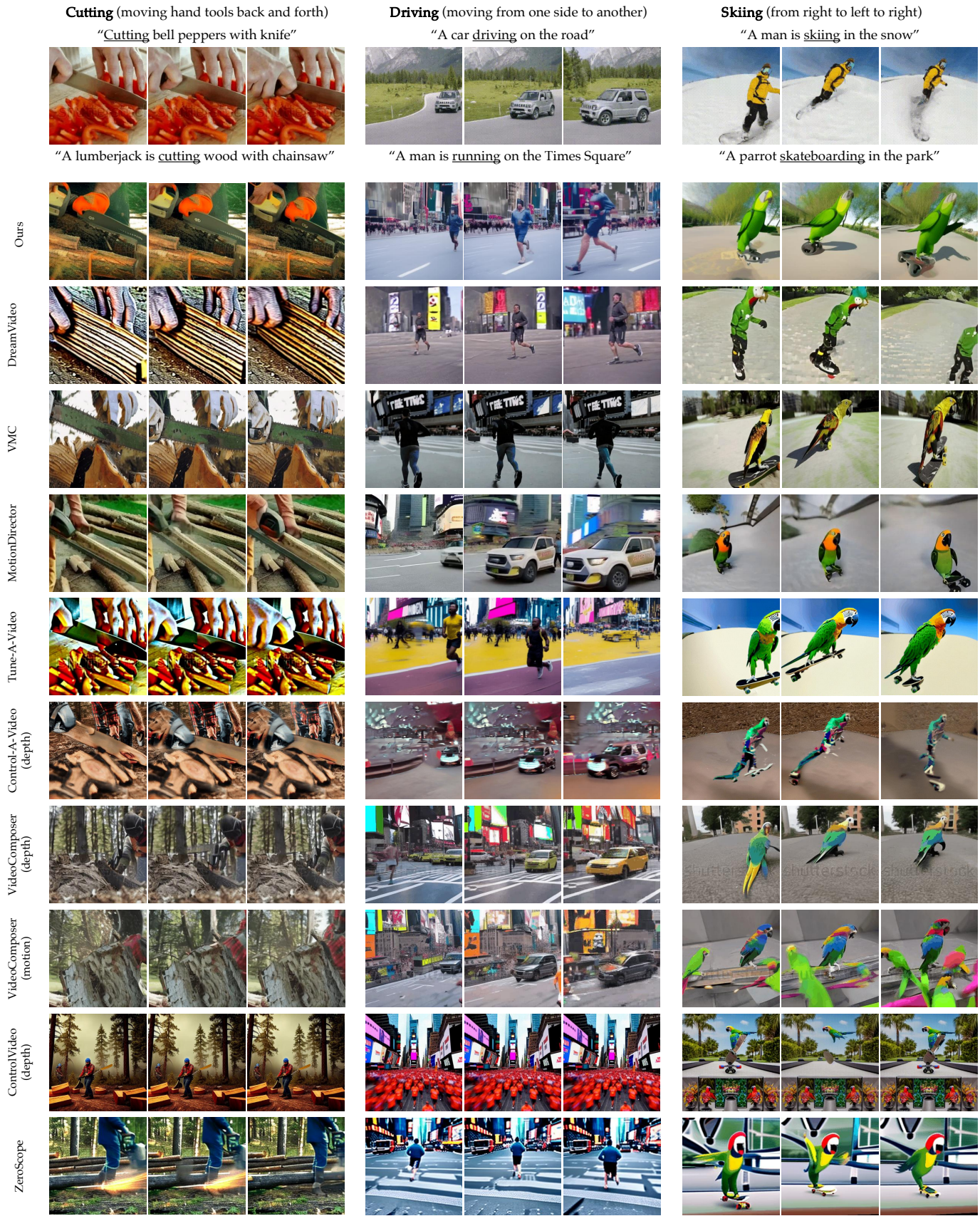



Fig. 4: Qualitative evaluation results. Our method outperforms state-of-the-art methods in appearance diversity and motion fidelity.

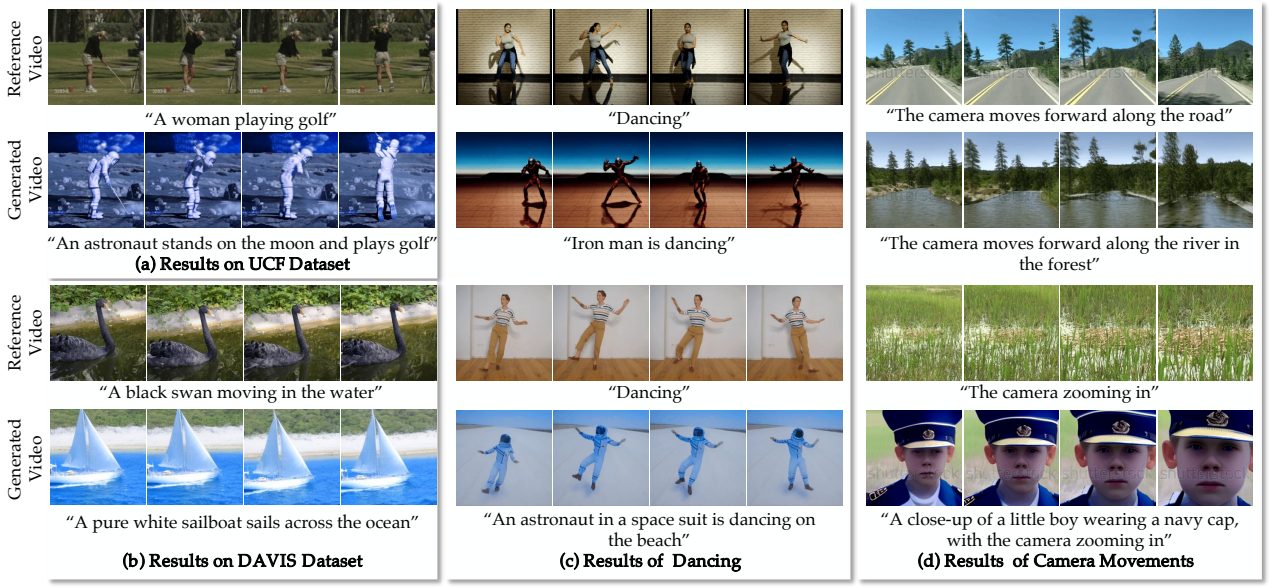



Fig. 5: Results of *MotionCrafter* in UCF dataset [37], DAVIS dataset [38], dancing videos and different camara movements.

motions. For each motion, we employ eight basic prompts that alter either the object or background, alongside eight more complex prompts that involve changes to both the object and background. Consequently, we obtained a total of 768 video clips for each method to compare.

**Implementation details:** In all photorealistic video generation experiments, we use ZeroScope [36] with the default network architecture. In all stylized video generation experiments, we use AnimateDiff [7] with the default network architecture. We set a learning rate of $2 \times 10^{-5}$. The resolution of the input video is $256 \times 256$ and each sequence contains 16 frames. The training process for each motion requires approximately $150 \sim 300$ iterations using an NVIDIA L40 with a batch size of 1. *MotionCrafter* dose not influence the inference time of the base model. The number of inference steps is set to $T = 25$ and the guidance scale is set to $w = 7.5$.

### 4.2 Qualitative Evaluations

As shown in Fig. 4, we conduct qualitative comparisons with nine state-of-the-art methods. To highlight *MotionCrafter*'s robust ability to decouple motion and appearance, we employ complex prompts that involve substantial content changes between the reference and generated videos, such as "knives" transforming into "chainsaws".

We choose the common used ZeroScope [36] as our baseline model. ZeroScope is capable of generating videos with the desired appearance. Due to the lack of motion control, it struggles to produce results that correspond to the target actions. DreamVideo [31] achieves motion customization by injecting an additional motion adapter and encodes the appearance image of the reference video with CLIP and broadcasts the embeddings during the training process to minimize fitting to the appearance during training. However, simply adding an extra appearance prompt without decoupling it with specific constraint leads to overfitting the appearance, as shown in Fig. 4. VMC [32] relys on motion vectors, and it encounters challenges in scenarios where the generated objects are inconsistent with the objects in the reference videos. As shown in Fig. 4, VMC fails to guide human motion using the movement of a vehicle. MotionDirector [30] struggles to effectively balance the removal of appearance features while preserving motion, particularly in complex scenarios such as transforming a car into a person, where motion characteristics may become coupled with the appearance of the reference video. Additionally, in cases involving intricate motions, such as the trajectory and posture associated with skiing, it tends to lose critical details. Tune-A-Video [18] is a fine-tuning-based method similar to ours, but its objective does not focus on decoupling motions. Therefore, it fails to handle significant changes in the appearance of the image and tends to generate results similar to the original video. We employ the depth-map control model of Control-A-Video [16] to minimize the impact of the original video's appearance. However, it fails to alter the shape of objects, generating a car in the second example and a distorted monkey in the third example. We use VideoComposer [19] with both depth-map control and motion control models. The depth-map control also faces challenges in altering the shapes of objects. VideoComposer represents video-specific elements using motion vectors, i.e., 2D vectors that capture pixel-wise movements between adjacent frames. However, this motion representation method fails to capture fine-grained motion patterns, such as the arm movement in the third example. We utilize the depth-map control model of ControlVideo [15]. The depth-wise constraint is relatively less restrictive, allowing drastic changes in the appearance to match the specified style. However, this method results in the loss of motion information from the reference video. *MotionCrafter* generates desirable content while ensuring consistent actions and high visual quality.

To demonstrate the generalization capability of *MotionCrafter*, as shown in Fig. 5 (a) and (b), we present the results in UCF Sports dataset and DAVIS dataset. *MotionCrafter* is able to achieve good results on different datasets and various types of sports. As shown in Fig. 5 (c), *MotionCrafter* can even

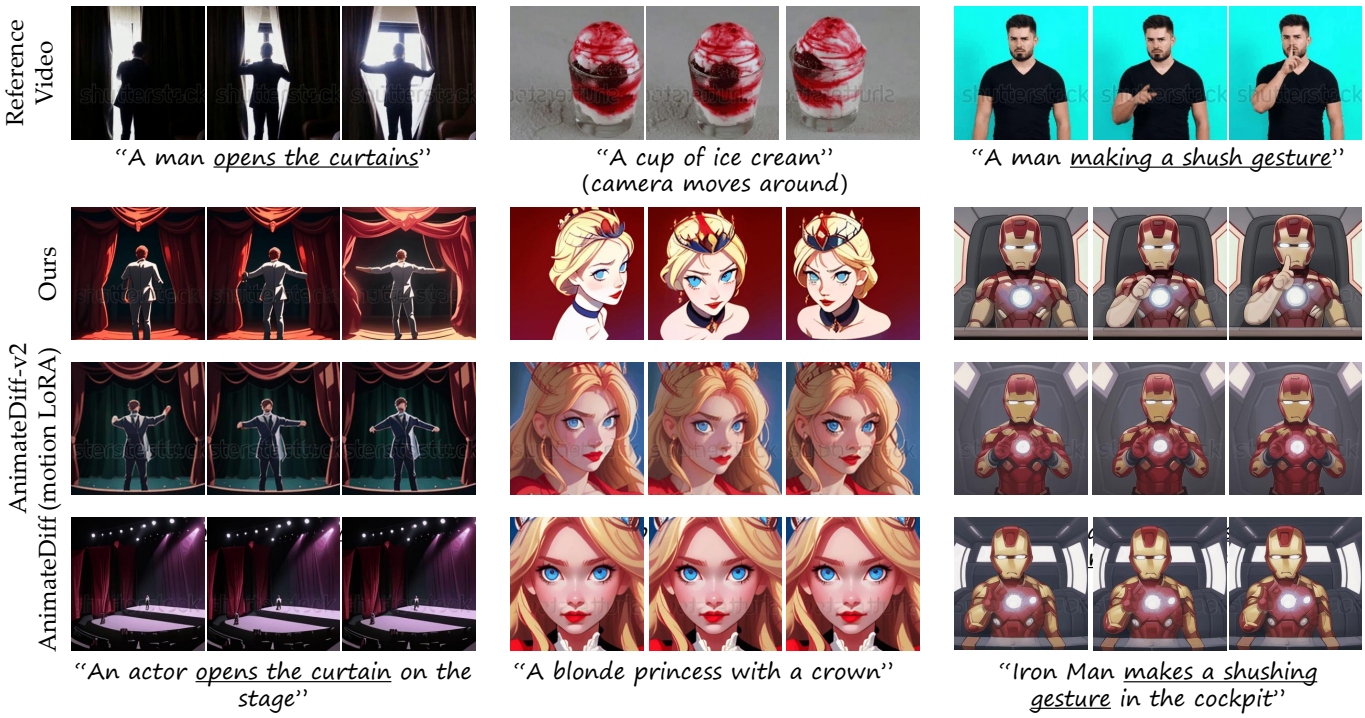



Fig. 6: Results of the integration of our *MotionCrafter* with AnimateDiff [7]. Without the guidance of *MotionCrafter*, it is difficult for the backbone model to generate obvious dynamics. Our method demonstrates strong generalization performance and can generate high-quality animation results with controllable motions.

TABLE 1: Quantitative evaluation results on UCF [37] dataset, DAVIS dataset [38] and our own dataset. The best numbers are in **bold** and the second best results are underlined.

| Method | Quantitative evaluation on UCF | | | | Quantitative evaluation on DAVIS | | | | Quantitative evaluation on own data | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Motion fidelity↑ | Diversity↑ | Consistency↑ | FID↓ | Motion fidelity↑ | Diversity↑ | Consistency↑ | FID↓ | Motion fidelity↑ | Diversity↑ | Consistency↑ | FID↓ |
| Zeroscope [36] | 0.6405 | 0.2462 | 0.8921 | 241.1 | 0.4405 | 0.2581 | 0.8854 | 238.1 | 0.4511 | 0.2465 | 0.8923 | 251.2 |
| Control-A-Video [16] | 0.6923 | 0.2123 | 0.7952 | 273.5 | 0.3915 | 0.2413 | 0.7742 | 252.5 | 0.3524 | 0.2126 | 0.7917 | 263.5 |
| ControlVideo [15] | 0.6125 | 0.2244 | 0.8273 | 278.6 | 0.4463 | 0.2138 | 0.8015 | 272.1 | 0.4223 | 0.2282 | 0.8571 | 278.6 |
| VideoComposer (depth) [19] | 0.7535 | 0.2416 | 0.8294 | 264.3 | 0.4721 | 0.2315 | 0.8354 | 249.7 | 0.4633 | 0.2424 | 0.8892 | 267.6 |
| VideoComposer (motion) [19] | 0.7748 | 0.2522 | 0.8862 | 275.1 | 0.4738 | 0.2682 | 0.8714 | 262.0 | 0.4746 | 0.2427 | **0.8980** | 270.4 |
| Tune-A-Video [18] | 0.7539 | 0.2215 | 0.8112 | **232.7** | 0.4322 | 0.2615 | 0.8321 | 229.4 | 0.4067 | 0.2209 | 0.8112 | **239.2** |
| MotionDirector [30] | 0.7514 | 0.2317 | 0.8756 | 235.7 | 0.6226 | 0.2678 | 0.8814 | 230.0 | 0.4522 | 0.2325 | 0.8854 | 255.7 |
| VMC [32] | 0.6724 | **0.2553** | 0.8955 | 265.7 | 0.6513 | 0.2647 | **0.8724** | 251.1 | 0.5252 | 0.2513 | 0.8854 | 252.1 |
| DreamVideo [31] | 0.7324 | 0.2413 | 0.8542 | 236.6 | 0.5842 | 0.2436 | 0.8613 | 240.4 | 0.4878 | 0.2213 | 0.8741 | 257.1 |
| Ours | **0.8514** | 0.2541 | **0.8958** | 235.5 | **0.6891** | **0.2698** | 0.8712 | **223.4** | **0.6502** | **0.2548** | 0.8956 | 250.6 |

generate complex dance movements. As shown in Fig. 5 (d), *MotionCrafter* can also maintain consistency in different camera motions.

We further validate the effectiveness of *MotionCrafter* on the basis of AnimateDiff and AnimateDiff v2 [7]. During inference, the motion module acquired by *MotionCrafter* can be combined with any desired generation backbone, resulting in the production of videos that exhibit controllable motions while possessing different stylistic effects. We select several representative personalized models contributed by artists affiliated with CivitAI [43]. As illustrated in Fig. 6, AnimateDiff v2 provides a fine-tunable motion module designed to achieve camera movements, such as the camera panning to the right depicted in the second column.). However, when it comes to more complex motions, the lack of strong constraints often makes it difficult to generate distinct dynamics. *MotionCrafter* is well-suited for extending the model from text-to-image to text-to-video generation.

### 4.3 Quantitative Evaluations

We measure the appearance diversity using the average CLIP [44] similarity between the diverse text prompts and all frames of the generated videos. We measure the temporal consistency using the average CLIP similarity between adjacent frames. We use the action classification method UniFormer [45] and calculate the average accuracy of each method. We use FID [46] score to evaluate the image quality. Since the FID score is computed between reference frames and generated frames, Tune-A-Video [18], which is spatially trained on the reference videos, get higher scores. Our method can injects motions into the backbone model without harming the visual quality. Table 1 presents the quantitative evaluation results of our method and the baseline approaches. We achieve state-of-the-art results in motion fidelity, diversity, and maintain good consistency and visual quality. VideoComposer [19] delivers desirable results in terms of diversity and consistency. However, as indicated by our user study results and qualitative comparisons, it may fall short of accurately transferring the target motion. VMC

TABLE 2: User study results. The best numbers are in **bold** and the second best results are underlined.

| Method | User preference | | | |
|---|---|---|---|---|
| | Motion fidelity↑ | Diversity↑ | Consistency↑ | Visual quality↑ |
| Zeroscope [36] | 2.05 | 2.57 | 2.70 | 2.46 |
| Control-A-Video [16] | 3.22 | 2.48 | 2.94 | 2.54 |
| ControlVideo [15] | 1.65 | 2.36 | 2.08 | 2.03 |
| VideoComposer (depth) [19] | 2.95 | 2.88 | 3.08 | 2.73 |
| VideoComposer (motion) [19] | 2.70 | 2.90 | 2.98 | 2.71 |
| Tune-A-Video [18] | 2.70 | 2.49 | 2.32 | 2.27 |
| MotionDirector [30] | 3.14 | 3.28 | 2.94 | 2.66 |
| VMC [32] | 4.12 | 3.70 | 3.04 | 4.04 |
| DreamVideo [31] | 3.32 | 3.61 | 2.49 | 3.26 |
| Ours | **4.34** | **4.33** | **4.16** | **4.12** |


Fig. 7: Ablation study results of *MotionCrafter*. Without U-Net layered tuning, the overall colors are more similar to the reference. Without timestep tuning, the accuracy of motion learning is lower. Without timestep-layered tuning, the magnitude of the motion is reduced. Without $\mathcal{L}_{appearance}$, the appearance and background are more similar to the reference.

adopts a multi-stage generation process, introducing a low-resolution motion generation model and a super-resolution model, thereby achieving better consistency.

### 4.4 User Study

We conduct a user preference assessment, comparing our approach with the aforementioned baseine methods. We employ a rating scale with four criteria: motion fidelity, appearance diversity, video consistency, and visual quality, applied to a dataset of 12 motions. In total, 102 participants took part in the survey. They were first informed about the objectives and settings of the motion customization task. Subsequently, we showed them reference videos, outputs from eight different methods, and the corresponding prompt conditions. The participants were asked to rate the outputs of each method using a five-point scale, with higher scores reflecting greater user satisfaction with the generated results. The user study results are presented in Table 2. Our method achieves the highest user preference, particularly in terms of motion fidelity and appearance diversity.

### 4.5 Ablation Study

We conduct ablation study to validate the effectiveness of the three key components of our method, i.e., parallel spatial-temporal architecture, dual-branch motion disentanglement and timestep-layered tuning.

**Ablation on extended T2I model.** The ablation study results on extended T2I model are presented in Fig. 7, along with a comparison to the baseline model AnimateDiff [7]. Due to the lack of inter-frame awareness in the generation backbone of extended T2I models, their dynamics are often more challenging to express than real T2V models, relying solely on an additional motion module. From the third row, it can be seen that incorporating timestep-related tuning significantly enhances the dynamic results and enables the learning of more accurate motion. However, without U-Net layered tuning, the overall appearance of the model's results is influenced by the reference video, leading to less vivid colors. The fourth row illustrates that models with U-Net layered tuning can generate more vivid and diverse appearances and artistic effects. However, without timestep-related tuning, there is a noticeable decrease in motion accuracy, and the overall dynamics are weakened. The fifth row shows the results of adding motion loss to the baseline, where the model can learn primary motion information. However, without time-layered tuning, the model struggles to learn stronger motion and the generated artistic effects are compromised. The last row shows the results of ablating time-layered tuning and motion loss, i.e., the baseline method. It can be observed that without dynamic guidance, the baseline model can generate very simple motions. The second row shows the results of the full model, which combines both motion accuracy and appearance diversity while maintaining the personalized generation backbone's artistic effects. Compared to these alternative baselines, our full model produces superior results, particularly in motion fidelity and appearance diversity.

**Ablation on T2V model.** The ablation results on the T2V model are illustrated in Fig. 8. In the third row, it is evident that without employing $\mathcal{L}_{appearance}$ to decouple appearance from motion, the model reproduces similar appearance features from the reference video. Furthermore, the model's capacity to learn motion is weakened, as shown in the fourth example, where the dynamic effect is absent in the output. In the fourth row, the generated video overfits to the reference video without separate tuning.

**Ablation on the hyper-parameter $\beta$.** As illustrated in Fig. 9, we present the ablation results for the hyper-parameter $\beta$ in the motion loss, which controls the strength of the appearance normalization loss. A large $\beta$ value leads to greater diversity in the generated appearance, while a small value results in an appearance closer to the reference video. We set $\beta$ to 1, 5, and 20 to evaluate its impact on the final results. When $\beta$ is set to 1, as shown in the second row, the model generates diverse content, but the shape of objects is difficult to alter. When $\beta$ is set to 5, as shown in the third row, the model produces diverse content with varied shapes, and the visual quality is satisfactory. However, when $\beta$ is set to 20, as shown in the fourth row, the generated videos exhibit blurriness and artifacts, resulting in lower video quality. Excessively large appearance normalization loss impairs the model's ability to learn appropriate motions and appearances from the reference and normalization videos. Consequently, we set $\beta = 5$ in all experiments shown in our main manuscript.

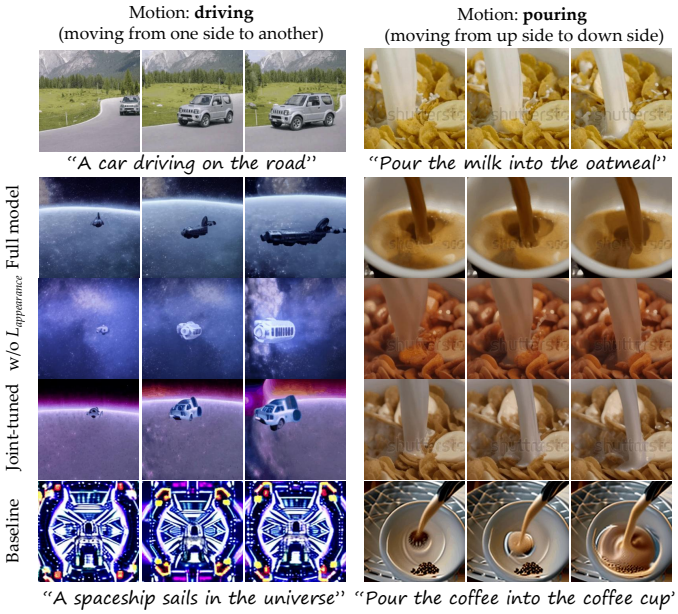



Fig. 8: Ablation study results of *MotionCrafter* on text-to-video generation backbone.

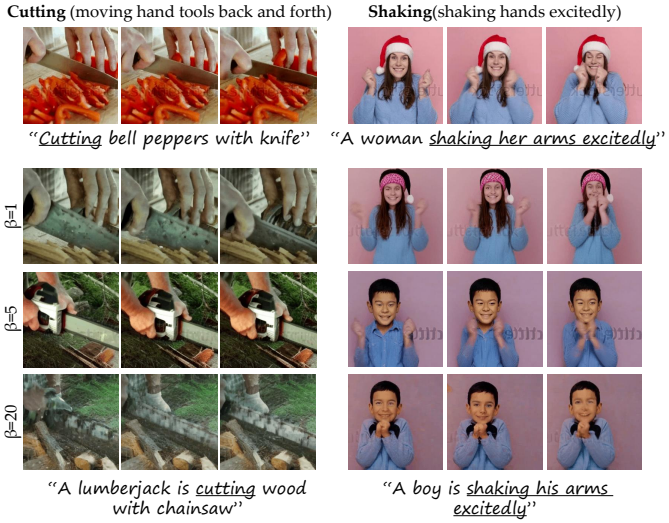



Fig. 9: Ablation study results of the hyper-parameter $\beta$ in our dual-branch motion disentanglement.

## 5 Conclusion

In this work, we tackle the challenge of motion customization in video generation by proposing a plug-and-play instance-guided approach. The proposed parallel spatial-temporal architecture can effectively separate motion and appearance, enabling the injection of reference motions into the temporal module of the base model. Additionally, our novel dual-branch motion disentanglement method successfully decouples appearance and motion, by incorporating a motion disentanglement loss and an appearance prior enhancement scheme. By proposing a timestep-layered tuning strategy, the efficiency and quality of motion learning can obviously be improved. The extensive quantitative and qualitative evaluations, as well as the user preference survey, demonstrate the effectiveness of *MotionCrafter*. *MotionCrafter*

can produce videos with a variety of artistic styles with various personalized backbones in the community. Our work paves the way for more motion-aware text-to-video generation methods.

**Limitations and Future Work** While we have demonstrated the ability of *MotionCrafter* to generate complex dynamic motions, there are certain limitations imposed by the model structure and computational resources. In cases of complex actions requiring extended durations for completion, such as a series of aerobics exercises, maintaining action coherence often necessitates learning from more than 24 frames. One potential approach to tackling this issue is segmenting sequential actions into multiple units or implementing interpolation between frames. Furthermore, for complex actions involving a group of individuals, such as scenes from a ballet group performance, *MotionCrafter* may struggle to accurately capture the detailed dynamics of each individual. This challenge arises partly from the intrinsic complexity of motions and is also due to the limitations of current text-to-video generation models in producing high-quality representations of group objects. Addressing the aforementioned challenges will be targeted in our future work.

## REFERENCES


[1] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22 500–22 510.

[2] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *International Conference on Learning Representations (ICLR)*, 2022.

[3] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 15 954–15 964.

[4] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan, "VideoFusion: Decomposed diffusion models for high-quality video generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 10 209–10 218.

[5] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *International Conference on Learning Representations (ICLR)*, 2023.

[6] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-Concept customization of text-to-image diffusion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 1931–1941.

[7] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin, and B. Dai, "AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning," in *International Conference on Learning Representations (ICLR)*, 2024.

[8] Y. He, M. Xia, H. Chen, X. Cun, Y. Gong, J. Xing, Y. Zhang, X. Wang, C. Weng, Y. Shan *et al.*, "Animate-A-Story: Storytelling with retrieval-augmented video generation," *arXiv preprint arXiv:2307.06940*, 2023.

[9] Y. Gong, Y. Pang, X. Cun, M. Xia, Y. He, H. Chen, L. Wang, Y. Zhang, X. Wang, Y. Shan, and Y. Yang, "TaleCrafter: Interactive story visualization with multiple characters," in *ACM SIGGRAPH Asia Conference Proceedings*, 2023, pp. 101:1–101:10.

[10] N. Huang, Y. Zhang, and W. Dong, "Style-A-Video: Agile diffusion for arbitrary text-based video style transfer," *IEEE Signal Processing Letters*, vol. 31, pp. 1494–1498, 2024.

[11] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, C. Weng, and Y. Shan, "VideoCrafter1: Open diffusion models for high-quality video generation," *arXiv preprint arXiv:2310.19512*, 2023.

[12] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen Video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[13] W. Wang, H. Yang, Z. Tuo, H. He, J. Zhu, J. Fu, and J. Liu, "VideoFactory: Swap attention in spatiotemporal diffusions for text-to-video generation," *arXiv preprint arXiv:2305.10874*, 2023.

[14] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "ModelScope text-to-video technical report," *arXiv preprint arXiv:2308.06571*, 2023.

[15] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian, "ControlVideo: Training-free controllable text-to-video generation," in *International Conference on Learning Representations (ICLR)*, 2024.

[16] W. Chen, J. Wu, P. Xie, H. Wu, J. Li, X. Xia, X. Xiao, and L. Lin, "Control-A-Video: Controllable text-to-video generation with diffusion models," in *International Conference on Learning Representations (ICLR)*, 2024.

[17] S. Yang, Y. Zhou, Z. Liu, and C. C. Loy, "Rerender A Video: Zero-shot text-guided video-to-video translation," in *ACM SIGGRAPH Asia Conference Proceedings*, 2023, pp. 95:1–95:11.

[18] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-Video: One-shot tuning of image diffusion models for text-to-video generation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 7623–7633.

[19] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou, "VideoComposer: Compositional video synthesis with motion controllability," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[20] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, "Make-a-Video: Text-to-video generation without text-video data," in *International Conference on Learning Representations (ICLR)*, 2023.

[21] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22 563–22 575.

[22] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[23] O. Avrahami, K. Aberman, O. Fried, D. Cohen-Or, and D. Lischinski, "Break-A-Scene: Extracting multiple concepts from a single image," in *ACM SIGGRAPH Asia Conference Proceedings*, 2023, pp. 96:1–96:12.

[24] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 10 146–10 156.

[25] Z. Huang, T. Wu, Y. Jiang, K. C. Chan, and Z. Liu, "ReVersion: Diffusion-based relation inversion from images," in *SIGGRAPH Asia 2024 Conference Papers*. New York, NY, USA: Association for Computing Machinery, 2024, pp. 4:1–4:11.

[26] Y. Wen, N. Jain, J. Kirchenbauer, M. Goldblum, J. Geiping, and T. Goldstein, "Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[27] A. Voynov, Q. Chu, D. Cohen-Or, and K. Aberman, "$p+$: Extended textual conditioning in text-to-image generation," *arXiv preprint arXiv:2303.09522*, 2023.

[28] Y. Zhang, W. Dong, F. Tang, N. Huang, H. Huang, C. Ma, T.-Y. Lee, O. Deussen, and C. Xu, "ProSpect: Prompt spectrum for attribute-aware personalization of diffusion models," *ACM Transactions on Graphics*, vol. 42, no. 6, pp. 244:1–244:14, 2023.

[29] X. Chen, L. Huang, Y. Liu, Y. Shen, D. Zhao, and H. Zhao, "AnyDoor: Zero-shot object-level image customization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 6593–6602.

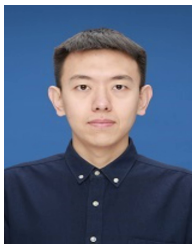
[30] R. Zhao, Y. Gu, J. Z. Wu, D. J. Zhang, J. Liu, W. Wu, J. Keppo, and M. Z. Shou, "MotionDirector: Motion customization of text-to-video diffusion models," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 273—290.

[31] Y. Wei, S. Zhang, Z. Qing, H. Yuan, Z. Liu, Y. Liu, Y. Zhang, J. Zhou, and H. Shan, "DreamVideo: Composing your dream videos with customized subject and motion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 6537–6549.

[32] H. Jeong, G. Y. Park, and J. C. Ye, "VMC: Video motion customization using temporal attention adaption for text-to-video diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 9212–9221.
[33] Y. Zhang, Z. Xing, Y. Zeng, Y. Fang, and K. Chen, "PIA: Your personalized image animator via plug-and-play modules in text-to-image models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 7747–7756.
[34] V. Jain, M. Rungta, Y. Zhuang, Y. Yu, Z. Wang, M. Gao, J. Skolnick, and C. Zhang, "HiGen: Hierarchy-aware sequence generation for hierarchical text classification," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 1354–1368.
[35] S. Yang, L. Hou, H. Huang, C. Ma, P. Wan, D. Zhang, X. Chen, and J. Liao, "Direct-a-Video: Customized video generation with user-directed camera movement and object motion," in *ACM SIGGRAPH 2024 Conference Papers*. New York, NY, USA: Association for Computing Machinery, 2024, pp. 113:1–113:12.
[36] "zeroscope_v2 576w," 2023, https://huggingface.co/cerspense/zeroscope_v2_576w. Last accessed on 2023-11-11.
[37] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Computer Vision in Sports, Springer International Publishing*, 2014.
[38] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 724–732.
[39] Y. Alaluf, E. Richardson, G. Metzer, and D. Cohen-Or, "A neural space-time representation for text-to-image personalization," *ACM Transactions on Graphics*, vol. 42, no. 6, pp. 243:1–243:10, 2023.
[40] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 22 560–22 570.
[41] Y. Frenkel, Y. Vinker, A. Shamir, and D. Cohen-Or, "Implicit style-content separation using B-LoRA," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 181—-198.
[42] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1728–1738.
[43] "Civitai," 2022, https://civitai.com/. Last accessed on 2023-11-11.
[44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.
[45] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "UniFormer: Unified transformer for efficient spatiotemporal representation learning," in *International Conference on Learning Representations (ICLR)*, 2023.
[46] M. Wright and B. Ommer, "ArtFID: Quantitative evaluation of neural style transfer," in *DAGM German Conference on Pattern Recognition*. Springer, 2022, pp. 560–576.

**Yuxin Zhang** received B.Sc. degree in Automation from Tsinghua University, Beijing, China, in 2020. She is now a Ph.D. candidate of the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, and the School of Artificial Intelligence at the University of Chinese Academy of Sciences. Her research interests include computer vision, computer graphics, and machine learning.

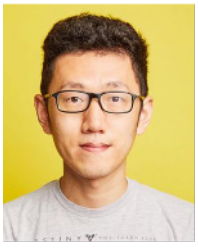
**Weiming Dong** (Member, IEEE) is a Professor at the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences. He received his BSc and MSc degrees in 2001 and 2004, both from Tsinghua University, China. He received his PhD in Computer Science from the Henri Poincaré University (Nancy 1), France, in 2007. His research interests include image synthesis, AIGC, and computational creativity.



**Fan Tang** received the B.Sc. degree in computer science from North China Electric Power University, Beijing, China, in 2013, and the Ph.D. degree from Institute of Automation, Chinese Academy of Sciences, Beijing, in 2019. He is an Associate Professor with the School of Computer Science and Technology, University of Chinese Academy of Sciences. His research interests include computer graphics, computer vision, and machine learning.

**Nisha Huang** received the B.S. degree in Aerospace Information Engineering from Beijing University of Aeronautics and Astronautics in 2021 and the M.S. degree in Artificial Intelligence from University of Chinese Academy of Sciences in 2024. She is currently pursuing her Ph.D. degree in Electronic Information at Tsinghua University. Her research interests include multimedia analytics, computer vision, and machine learning.

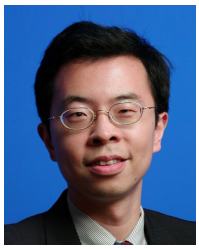
**Haibin Huang** received his BSc and MSc degrees in Mathematics in 2009 and 2011 respectively from Zhejiang University. He obtained his Ph.D. in Computer Science from UMass Amherst in 2017. He is currently a Senior Staff Research Scientist at Kuaishou Technology.



**Chongyang Ma** received B.S. degree from the Fundamental Science Class (Mathematics and Physics) of Tsinghua University in 2007 and Ph.D. degree in Computer Science from the Institute for Advanced Study of Tsinghua University in 2012. He is currently a Research Lead with Kuaishou Technology, Beijing. His research interests include computer graphics and computer vision.

**Pengfei Wan** received the BE degree in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, and the PhD degree in electronic and computer engineering from the Hong Kong University of Science and Technology, Hong Kong. He is the head of Visual Generation and Interaction Center, Kuaishou Technology. His research interests include image/video signal processing, computational photography, and computer vision.

**Tong-Yee Lee** (Senior Member, IEEE) received the PhD degree in computer engineering from Washington State University, Pullman, in May 1995. He is currently a chair professor with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan. He leads the Computer Graphics Laboratory, National Cheng-Kung University (http://graphics.csie.ncku.edu.tw). His current research interests include computer graphics, nonphotorealistic rendering, medical visualization, virtual reality, and media resizing. He is a senior member of the IEEE Computer Society and a member of the ACM. He also serves on the editorial boards of the IEEE Transactions on Visualization and Computer Graphics.



**Changsheng Xu** (Fellow, IEEE) is a Professor at the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences. Dr. Xu received the Best Associate Editor Award of ACM Transactions on Multimedia Computing, Communications and Applications in 2012 and the Best Editorial Member Award of ACM/Springer Multimedia Systems Journal in 2008. He has served as an Associate Editor, a Guest Editor, a General Chair, a Program Chair, an Area/Track Chair, a Special Session Organizer, a Session Chair, and a Transactions on Professional Communication (TPC) Member for over 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops. He is the International Association for Pattern Recognition (IAPR) Fellow and the ACM Distinguished Scientist.