Mobius: Text to Seamless Looping Video Generation via Latent Shift

XIULI BI, Chongqing University of Post and Telecommunications, China

JIANFEI YUAN, Chongqing University of Post and Telecommunications, China

BO LIU, Chongqing University of Post and Telecommunications, China

YONG ZHANG, Meituan, China

XIAODONG CUN*, GVC Lab, Great Bay University, China

CHI-MAN PUN, University of Macau, China

BIN XIAO, Chongqing University of Post and Telecommunications, China

Prompt: *"A young female activist stands tall, holding a flag high above her head with determination in her eyes. The flag flutters in the breeze, its bold colors contrasting with the backdrop of a city street or public space. Her posture is confident, embodying strength."*

Prompt: *"A sleepy koala, nestled comfortably on a tree branch, lazily munches on eucalyptus leaves, its fluffy grey fur blending with the textured bark of the tree. The leaves sway slightly in the breeze as the koala picks them one by one, its black nose twitching with each bite."*

Fig. 1. Without any training, the proposed Mobius can generate seamless looping videos using the pre-trained Text-to-Video latent diffusion model directly. Can you identify the end in the above video? *Best viewed with Acrobat Reader. Click the video to play the animation clips.* **We also give these examples in the supplementary video.** Project page: http://mobius-diffusion.github.io.

We present Mobius, a novel method to generate seamlessly looping videos from text descriptions directly without any user annotations, thereby creating new visual materials for the multi-media presentation. Our method repurposes the pre-trained video latent diffusion model for generating looping videos from text prompts without any training. During inference, we first construct a latent cycle by connecting the starting and ending noise of the videos. Given that the temporal consistency can be maintained by the context of the video diffusion model, we perform multi-frame latent denoising by gradually shifting the first-frame latent to the end in each step. As a result, the denoising context varies in each step while maintaining consistency throughout the inference process. Moreover, the latent cycle in our method can be of any length. This extends our latent-shifting approach to generate seamless looping videos beyond the scope of the video diffusion model's context. Unlike previous cinemagraphs, the proposed method does not require an image as appearance, which will restrict the motions of the generated results. Instead, our method can produce more dynamic motion and better visual quality. We conduct multiple experiments and comparisons to verify the effectiveness of the proposed method, demonstrating its efficacy in different scenarios. All the code will be made available.

1 INTRODUCTION

Looping video, also called *cinemagraph* in some research, aims to create a seamless looping video without ends via periodical motions. It is a unique way to share a specific moment's dynamics, which is popular as short videos and animated GIFs on social media, photo-sharing platforms, and screen savers[1] to create a better user experience. However, capturing these looping videos needs huge manual efforts, including the stabilization of the camera, manually annotating the moving object, selecting the animated frames, *etc*.

Previous efforts [Bai et al. 2013; Halperin et al. 2021; Holynski et al. 2021; Liao et al. 2013] make cinemagraphs from the given video or a single image animation. However, due to the difficulty of modeling open-world motion prior, these methods only focus on creating

*Corresponding Author

Authors' addresses: Xiuli Bi, Chongqing University of Post and Telecommunications, Chongqing, China; Jianfei Yuan, Chongqing University of Post and Telecommunications, Chongqing, China; Bo Liu, Chongqing University of Post and Telecommunications, Chongqing, China; Yong Zhang, Meituan, Shenzhen, China; Xiaodong Cun, GVC Lab, Great Bay University, Dongguan, China, cun@gbu.edu.cn; Chi-Man Pun, University of Macau, Macau, China; Bin Xiao, Chongqing University of Post and Telecommunications, Chongqing, China.

[1] cinemagraphs.com

the looping video on the specific kinds, for example, water [Holyn-ski et al. 2021; Liao et al. 2013; Mahapatra et al. 2023], periodic pattern [Halperin et al. 2021], portrait [Bai et al. 2013; Bertiche et al. 2023; Zhang et al. 2022], panoramic [Agarwala et al. 2005; He et al. 2017]. Since the diffusion model provided universal genera-tive priors for video, current frame interpolation methods [Wang et al. 2024a,c] can naturally produce the cinemagraph by setting the same beginning and end frames, however, the generated results in frame interpolation will often tend to generate still results in all frames. Besides, all current cinemagraph methods focus on simple motions with limited movement, whereas real-world videos are more complex.

We define a new research problem beyond current cinemagraph synthesis which is directly generating the seamless looping video from text description. Different from previous methods which need tricks of a stable camera and only repeating some of the elements, our method aims to generate fully looping videos directly from the pre-trained text-to-video models, which will show more dynamic motions and natural visual effects, including the moving objects, the camera, *etc.*, by the generative prior. This is fully automatic and can generate videos which is unusual in real life. However, there are two key challenges in adapting it for our task. First, as the text-to-video diffusion model is trained on natural video, it remains unclear how to adapt it to our looping video generation. On the other hand, the current text-to-video generation model can only generate certain frames during inference. However, a short video might not provide a good representation of real-world dynamics.

Thus, we present Mobius to solve these problems in a training-free manner, where the key observation is that each frame should be considered equally important in the video final video. To this end, firstly, we propose a latent shift strategy in denosing. We con-struct a cycle utilizing all the noisy latents from the first frame to the end frame. Then, we shift its position by adding the first frame to the last to build the new noisy latent for denoising. Thus, the video model maintains temporal consistency in each denoising step, and each video is equally considered. For the generation in the longer context, the proposed latent shifting strategy naturally enables longer looping video generation by a longer denoising se-quence cycle. However, if we directly generate the longer video utilizing this method, the generated results are also influenced by the inaccurate position embedding and frame-variant 3D VAE. Thus, we extend the rotary position embedding by an NTK-aware in-terpolation method inspired by the long context Large Language Model [Peng et al. 2023] and propose a frame-invariance method for latent decoding. Based on these modifications, the proposed method can directly utilize the pre-trained video diffusion model to generate high-quality cinemagraphs from text descriptions. Besides, we also show that the proposed latent shift can also work well for longer video generation tasks. Finally, the experiments demonstrate the qualitative and quantitative advantages of our approach.

Overall, the contribution can be summarized as:

- We conduct a new research problem for the open-domain seamless looping video generation from text description using a pre-trained text-to-video diffusion model.

- We propose a latent shifting strategy to interactively denoise the latent in each step so that we can generate the looping video and it can be in arbitrary lengths.
- The detailed experiments show that the proposed method can achieve state-of-the-art performance on looping video generation and we also give the applications on longer video generation.

2 RELATED WORK

2.1 Cinemagraphs and Looped Video Generation

Our task is similar to cinemagraphs, which aim to produce loop-ing videos by manipulating an input video manually. However, the manual creation of cinemagraphs is a time-consuming process, even for professional artists. Previously, learning-based methods faced difficulties in generating or editing an entire video. As a re-sult, prior techniques only applied to specific patterns to create cinemagraphs, for example, water [Holynski et al. 2021; Liao et al. 2013; Mahapatra et al. 2023], periodic pattern [Halperin et al. 2021], portrait [Bai et al. 2013; Bertiche et al. 2023; Zhang et al. 2022], panoramic [Agarwala et al. 2005; He et al. 2017]. As for represen-tative work, Endless Loops [Halperin et al. 2021] utilizes CRF to compute loop shifts, and it can only work on the repeated pattern. [Holynski et al. 2021] presents an image animation method to gener-ate the moving water from a single image utilizing Eulerian motion fields. Text-to-cinemagraphs [Mahapatra et al. 2023] further extend it by the pre-trained text-to-image stable diffusion model. Several methods [Li et al. 2024; Niu et al. 2024; Shi et al. 2024] present a two-stage framework to generate the video with trajectory control. However, they only work on certain object types or need manual trajectory design. While LoopAnimate [Wang et al. 2024a] employs multi-stage training and symmetric guidance to achieve looped generation, their generated results are too still. Besides, we can naively utilize the generative frame interpolation methods [Wang et al. 2024c; Xing et al. 2024] based on the video diffusion model for a generation by setting the same start and end keyframes. How-ever, since the original frame interpolation model is not trained for cinemagraph, the generated results might also be still. Besides, these methods involve additional larger-scale training for genera-tion, which might cause forgetting problems. Differently, we directly generate the looped video from the text description, yet with better visual effects, such as the whole movement of the camera and object motion.

2.2 Video Generation in Diffusion Model Era

Due to the stabilizing training process of the Diffusion Model [Ho et al. 2020; Song et al. 2020], video generation has had a big break-through in recent years. Eary works [Ho et al. 2022; Singer et al. 2022] directly generate high-resolution videos from cascade models of spatial and temporal layers directly in pixel space. On the other hand, utilizing the pre-trained text-to-image models [Rombach et al. 2022], *i.e.*, Stable Diffusion, as the base model, many works try to add additional layers to keep temporal consistency. [Chen et al. 2023; Wang et al. 2023b, 2024b] add temporal attention modules to the base model and train in an end-to-end fashion. Besides, [Guo et al. 2023] finds that training the models by temporal layers only
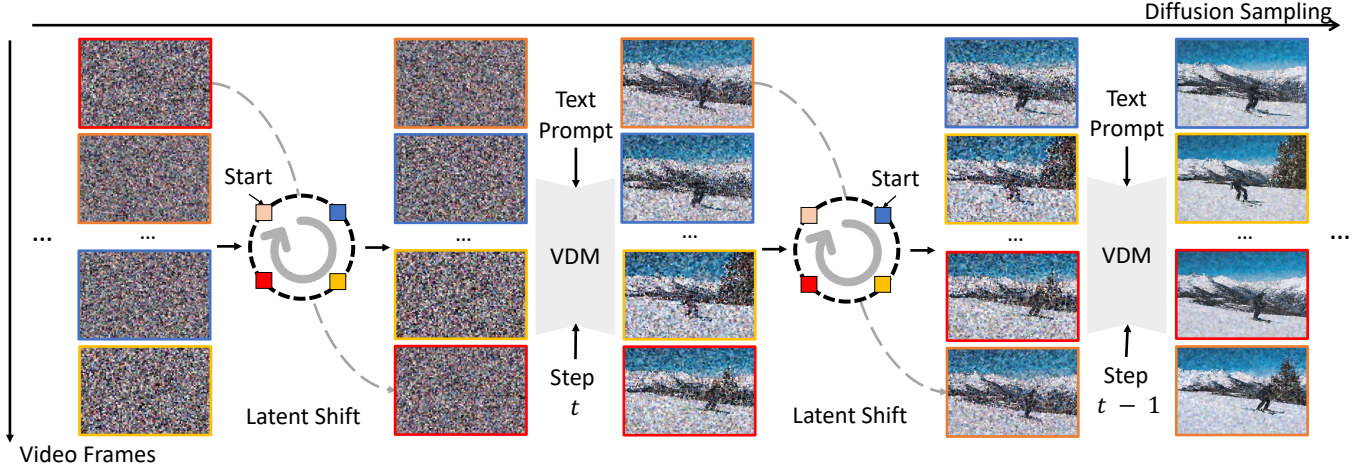
Fig. 2. **Latent Shift for looping video generation**. Taking 4 latent toys pre-trained Video Diffusion Models (VDM) as an example, we build a latent cycle and shift the start point in each denoising step in inference for text-guided looping video generation. Notice that, the shifting is conducted in the latent space, we emit the latent encoder and decoder for easy understanding.

has a better visual quality. [Chen et al. 2024] proposes a method to increase the visual qualities by a two-stage image and video joint training process. However, these methods only create a short video with limited motions, which restricts its applications in real-world cases. Besides the text-to-video diffusion model, new works also train image-to-video models for generation, which is also related to our task. For example, Stable Video Diffusion [Blattmann et al. 2023] fine-tunes the text-to-video diffusion model with a high-quality data pipeline. DynamicCrafter [Xing et al. 2025] shares a similar idea and trains on the video diffusion model. ToonCrafter [Xing et al. 2024] and Generative image in-between [Wang et al. 2024c] are further finetuning the image-to-video models for the generative frame interpolation. However, as we discussed before, directing utilizing the frame interpolation methods for our task might have issues with the too-still motion. Recently, Sora [Brooks et al. 2024] has made a big step in video generation via denoising transformers (DiT [Peebles and Xie 2023]), showing the scalability and advantages. Thus, the more recent video generation methods [Team 2024a,b; Yang et al. 2024] are based on the DiT structure, which has better motion and temporal consistency than previous methods.

Besides, since these pre-trained large diffusion models are trained from larger-scale datasets, we can repurpose these models for the new task without training. For instance, in the field of image/video editing, works such as Prompt to Prompt [Hertz et al. 2022], FateZero [Qi et al. 2023], and MasaCtrl [Cao et al. 2023] have achieved zero-shot editing through attention control. Meanwhile, there also contains some methods that have provided foundational discoveries for zero-shot editing [Yu et al. 2023a,b] and improving the performance without additional training [Si et al. 2024; Wu et al. 2023]. In this paper, we utilize the most popular open-sourced DiT-based video generation model, *i.e.*, CogVideoX [Yang et al. 2024], as the base model for looped video generation in a training-free manner.

2.3 Longer Video Generation in Diffusion Model

Our looping videos can be considered an infinitely longer video generation. In current methods, due to the limited latent length

in training the pre-trained text-to-video generation models, several methods are proposed to modify the denoising process of the original diffusion model for new purposes. For the longer video generation, Gen-L-Video [Wang et al. 2023a] uses the weighted sum of different short latent segments in the overlapping area to alleviate the inter-frame continuity issue. However, this method significantly increases the inference time and can lead to smooth transitions between frames. FreeNoise [Qiu et al. 2023] introduces a shuffled latent sequence design and uses attention-based weighting to maintain visual consistency in long videos. However, since the latent changes only occur in the shuffling, the resulting video motion may appear too static and is prone to out-of-memory (OOM) errors. FIFO [Kim et al. 2024] uses diagonal denoising for long video generation, maintaining the consistency and coherence of the video. However, there is a training inference gap for reasoning at different noise levels, and it lacks global information modeling. Video-Infinity [Tan et al. 2024] uses distributed inference to facilitate global and local information interaction, achieving video consistency while accelerating inference. However, an important limitation is the need for multiple GPUs to run simultaneously, and the quality of generating longer videos is not very good. DiTCtrl [Cai et al. 2024] utilizes a mask-based attention-sharing mechanism to maintain semantics, as well as a latent mixing strategy to achieve smooth transitions between video frames. However, this also brings about a significant amount of additional computational costs. These longer video generation methods change the combination of latent in the test time to control the generated content in the diffusion process, which inspired our looping video generation from text directly.

3 METHOD

Given the text prompt, we design a training-free method for generating the looping video by shifting the noise in each inference step of the pre-trained video diffusion model, so that all the frames will be considered equally in the final generated video. Below, we first introduce the basic paradigm of text-to-video to better understand our method in Sec. 3.1. Then, we introduce the proposed *Latent*

Shifting, which iteratively transforms the position of the latent in each step (Sec. 3.2). Since directly utilizing the looping latents will show artifacts when 3D VAE decoding, we design a frame-invariant decoding to decode looping video (Sec. 3.3). Finally, we introduce the *Rotary Position Encoding interpolation* to model global positional information when generating the longer looping videos in Sec. 3.4 and give some applications in Sec. 3.5, respectively.

3.1 Preliminary: Text-to-Video Latent Diffusion Model

Taking one specific video diffusion model, *i.e.*, CogVideoX [Yang et al. 2024], as an example, we introduce the basic concepts and knowledge of the text-to-video latent diffusion model. Current large text-to-video models are all based on the latent diffusion model [Rombach et al. 2022]. The latent diffusion model contains an auto-encoder $\mathcal{E}(\cdot)$, $\mathcal{D}(\cdot)$ for compressing the videos into the latent space. In the most advanced video diffusion models [Brooks et al. 2024; Team 2024a; Yang et al. 2024], the compression of videos in both the spatial and temporal domains represents the crucial factor for realizing better visual and temporal qualities. Then, following the Denoising Diffusion Probabilistic Models [Ho et al. 2020], for training, the input F frame video clip $v \in \mathbb{R}^{F \times H \times W \times 3}$ with width W and height H is first converted to the latent space $\mathbf{z}_0$, where $\mathbf{z}_0 = \mathcal{E}(v) = [z_0^1; ...; z_0^f] \in \mathbb{R}^{f \times h \times w \times c}$. h, w, f are the compressed height, width, and frame in the latent space, respectively. Then, the latent diffusion model ϵ_θ is trained to denoise its perturbed version $\mathbf{z}_t$. For noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, the time step of diffusion model $t \sim \mathcal{U}([1, ..., T])$, the text prompt c, this denoising diffusion model is trained to minimize the following loss:

$$\mathcal{L} = ||\epsilon - \epsilon_\theta(\mathbf{z}_t; c, t)||_2^2. \tag{1}$$

Here, the denoising network ϵ_θ is based on the DiT [Peebles and Xie 2023] architecture.

After training, giving any noise latents $[z_t^1; ...; z_t^f] \sim \mathcal{N}(0, \mathbf{I})$ for video generation, and a diffusion sampler $\Phi(\cdot)$, such as DDIM sampler [Song et al. 2020], the diffusion model generate the final clear video via an T-step iterative denoising, where t-th denoising step is expressed as:

$$[z_{t-1}^1; ...; z_{t-1}^f] = \Phi([z_t^1; ...; z_t^f], t, c; \epsilon_\theta), \tag{2}$$

where z_t^i denotes the latent of i-th frame at time step t. Notice that, the context length of the video diffusion model is restricted by the denoising network ϵ_θ, and each latent has the unchanged position when inference.

Finally, we could generate a video by the pre-trained latent decoder $\mathcal{D}(\cdot)$ of the 3D VAE as: $v' = \mathcal{D}(\mathbf{z}_0')$. Notice that, since the 3D VAE of the video diffusion model supports both image and video generation, they usually treat the first frame differently in temporal compression.

3.2 Latent Shifting

The text-to-video diffusion model is trained on a multi-frame latent diffusion model, where multiple latents are sent into the denoising network for a generation. Since our looping video requires each frame to be considered as the first frame, we thus need to make each latent have the temporal consistency of the previous latent and the

First frame   w/o Frame-Invariance decoding   Our last frame
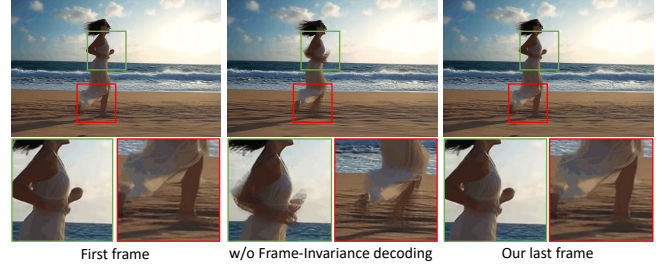
Fig. 3. Frame-invariance latent decoding reduces the artifacts caused by the 3D VAE decoding.

end latent. As shown in Figure. 2, we first build a cycle latent list for denoising by connecting the first frame latent and the last. Then, for each denoising step, we shift the first frame to the last to build a new multi-frame latents for generation. After multi-step denoising, we can maintain the whole temporal consistency of the entire video.

Formally, given the inference context f of a video diffusion model, we can generate the looping video which contains N latents, where $N = n \times f$ and n are the multiple factors for longer looping video generation. Firstly, we initialize all the latent as $[z_T^1; ...; z_T^N] \sim \mathcal{N}(0, \mathbf{I})$, then, for t-th denoising step, we shift the start point of the denoising context by $j = (t \times s) \mod N$, where s is the skip step of each iteration. Since we also need to maintain the f-frame inference restriction in the pre-trained diffusion model, the denoising step of this step can be formulated as:

$$[z_{t-1}^j; ...; z_{t-1}^{j+f-1}] = \Phi([z_t^j; ...; z_t^{j+f-1}], t, c; \epsilon_\theta), \tag{3}$$

where Φ is a DDIM Sampler [Song et al. 2020] as introduced before. When $j + f - 1 > N$, our cycle list creates the denoising latents by concat the $[z_t^j, ..., z_t^N]$ and $[z_t^1, ..., z_t^{f-(N-j+1)}]$.

Our latent shifting algorithm utilizes the multi-frame denoising steps in the diffusion model and the temporal consistency denoising of the video diffusion model for looping video generation. Notice that, since this latent denoising method can be any length, our method can produce any length inference looping videos and can also be utilized in the longer video generation.

3.3 Frame-Invariance Latent Decoding

To meet the demands of both text-to-video and text-to-image joint training, the latent compression of current state-of-the-art video generation models [Yang et al. 2024] does not compress each frame equally in the temporal dimension. In detail, CogVideoX employs a 3D VAE structure that compresses video frames both in spatial and temporal compression. However, the first latent frame employs a special encoding and does not do any compressions, while subsequent frames are encoded with the standard 4× compression for the motion similarity. In latent decoding, it utilizes the first three latent to generate the first night frames of video. This inconsistent treatment of latent is inherently incompatible with our proposed latent shift method for looping video generation since we aim to produce a looping video in which each frame should be considered equally. If we directly utilize the original 3DVAE, it results in artifacts in the generated first frame due to the 4× compression, as shown in Figure 6. To mitigate this issue, we copy the last three
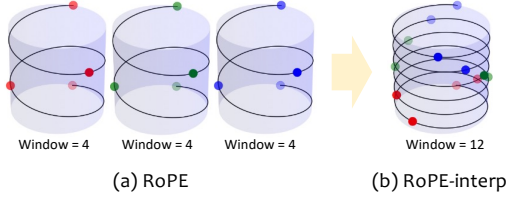
Fig. 4. We illustrate this with the example of the toy latent video diffusion model with a context window equal to 4. The utilized RoPE-Interp. enables longer video context without training by interpolation.

latents and insert them before the first latent as redundant frames to counteract the special compression of the first frame. Then, in the generated video, we remove the redundant generated frames by the added latent.

3.4 Rotary Position Embedding Interpolation

CogVideoX [Yang et al. 2024] uses Rotary Position Embedding (RoPE) to give positions in the attention model for denoising, which aims to achieve relative position encoding via absolute rotary position. However, if we directly utilize the original PoPE for our longer looping video generation task, the longer context does not match the original text-to-image model. To address this issue, we utilize a RoPE [Su et al. 2024] interpolation method for globally latent coding in the temporal dimension, inspired by the NTK-Aware interpolation in the longer context large language model [Peng et al. 2023].

Given the query vector at the m position q_m and the key vector at the n position k_n in the attention, RoPE introduces absolute positional information before calculating attention as follows:

$$
\begin{aligned}
Q_m &= RoPE(q_m, m) = q_m e^{im\theta}, \\
K_n &= RoPE(k_n, n) = k_n e^{in\theta}.
\end{aligned}
\tag{4}
$$

Here, $\theta = \text{diag}(\theta_0, \cdots, \theta_{d/2-1})$ is a pre-define diagonal matrix, where $\theta_i = b^{-2(b-1)/d}$, with $b = 10000$, and d represents the vector dimension. Then, we perform an inner product calculation to obtain the attention weights $A_{m,n}$ as follows:

$$
A_{m,n} = \text{Re}[\langle Q_m, K_n \rangle] = \text{Re}[\langle q_m, k_n \rangle \, e^{i(m-n)\theta}].
\tag{5}
$$

The result can be transformed into a value related to $m - n$, thus achieving relative position encoding.

To extend the encoding for longer lengths, we scale the base b as follows:

$$
b' = b \cdot k^{d/(d-2)}
\tag{6}
$$

Here, b' denotes the result after scaling, k represents the multiple by which the video length increases, and d indicates the dimension of the latents vector. Fig. 4 gives an illustration on how the RoPE-Interp. works. Since our core idea is to make each frame equal in the generation, we also try two different schemes to add the RoPE-Interp. to the features. The first one is the *shifted RoPE-Interp.*, where RoPE changes along with the latents, and another is the *fixed RoPE-Interp.*, where RoPE remains unchanged while the latents shift. We provide a more detailed comparison in the experiments.

3.5 More Application: Longer Video Generation

Longer video generation is an active research topic in current video generation since current text-to-video generation methods can only generate videos with limited context. The proposed latent shift naturally supports longer video inference beyond the training context by a non-cycle latent displacement. We utilize the same RoPE interpolation as we introduced before to correct the position of the latent. We give some examples in the supplemental videos.

4 EXPERIMENT

4.1 Settings and Implement Details

Implementation details. Our method is based on the pre-trained state-of-the-art open-source latent video diffusion model, CogVideoX-5B [Yang et al. 2024]. Notice that, we only modify the latent input of the diffusion model, our method might also work on any newly designed text-to-video latent diffusion models [Team 2024a,b], without training. Each video has a resolution of 480x720, and the inference step is set to 50 following a standard DDIM sampling strategy. Other parameters are the same as the default settings of CogVideoX. To evaluate the proposed methods, we choose 140 prompts from VBench [Huang et al. 2024] and EvalCrafter [Liu et al. 2024] and use GPT [Liu et al. 2023] to expand them into more detailed descriptions. All the experiments are conducted on a single NVIDIA H100 GPU. Since we only add a temporal latent shift in each step denoising, the proposed method has a similar inference speed compared with direct generation.

Baseline. Since there is no previous work for open-domain looping video generation from a text description, we majorly compare two generative interpolation methods and one method from the community. The first generative interpolation method is *Svd-Interp.* from Generative Image Inbetween [Wang et al. 2024c], which is trained on the stable video diffusion model [Blattmann et al. 2023] for frame interpolation. The other generative interpolation is *CogX-Interp.*[2], which is also trained from the image-to-video model of the CogVideoX for frame interpolation. To compare, we consider the first frame of our generated results for the starting and ending key frames of the interpolation. Notice that these two methods are based on larger-scale training for frame interpolation. Our method generates the looping video from the text description directly. *Latent Mix* is a method to achieve this looping video, which has been reported on Github[3], we compare with this method directly.

Evaluation Metrics. We report the MSE of the first frame and the last frame in the generated videos due to the looping video has the same first frame and the end frame. For the overall video quality, we utilize the widely used FVD [Unterthiner et al. 2018] and CLIP Score [Radford et al. 2021] for comparison. Besides, we give the overall video smoothness and dynamic score of the whole video from VBench [Huang et al. 2024].

4.2 Comparison with Other Methods

As introduced before, since current cinemagraph methods can not work on open-domain looping video generation, we compare our

---
[2]https://github.com/feizc/CogvideX-Interpolation
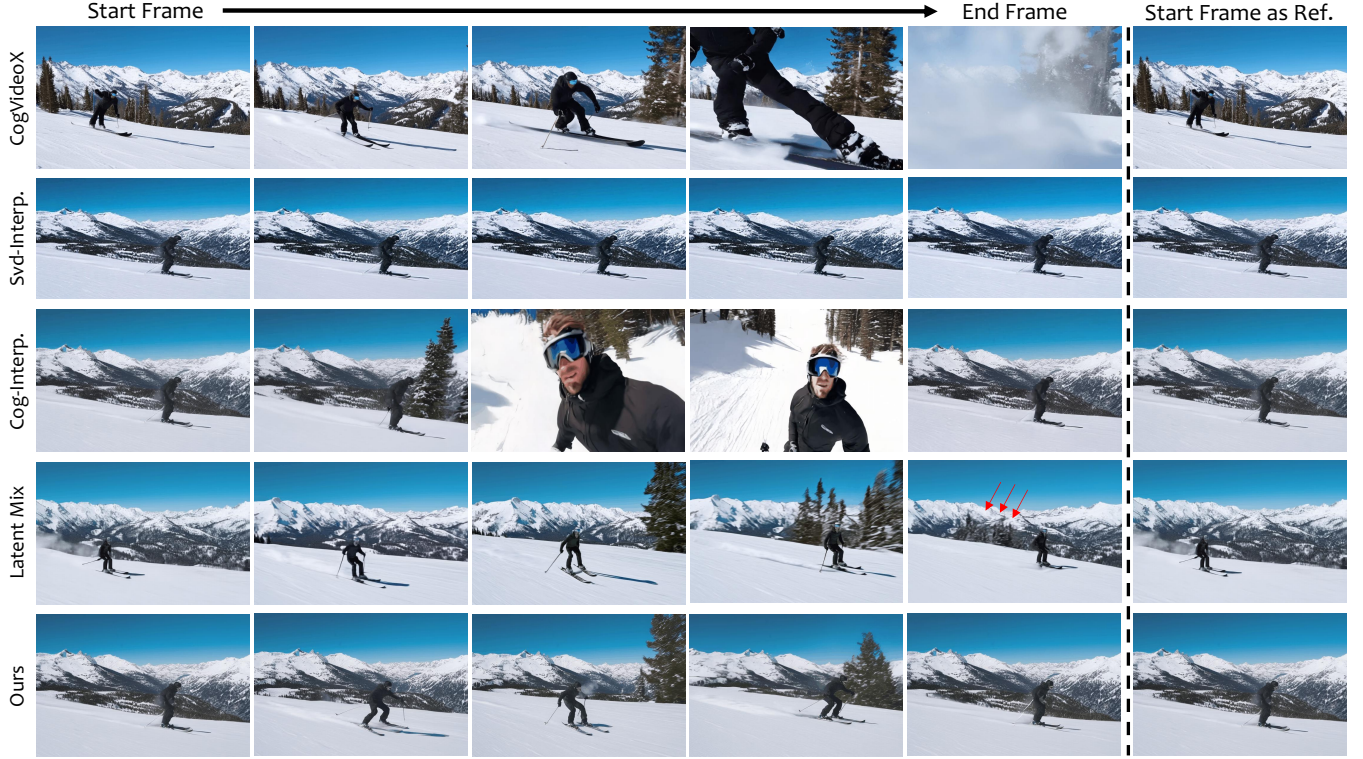[3]https://github.com/THUDM/CogVideo/issues/149

Fig. 5. Compare with other methods. We give the first frame, the intermediate frame, and the last frame for comparison. Notice that, both Svd-Interp. and Cog-Interp. are frame-interpolation methods, we manually give the same start frame and end frame as key-frames.

Table 1. Quantitative experimental results for different methods under the numerical evaluation metrics. * for the interpolation-based method, we utilize our generated first frame for the start and end keyframe, thus the MSE between the two frames is the oracle value.

	MSE↓	FVD↓	CLIP↑	Motion Smooth↑	Dynamic Score↑
Svd-Interp.*	18.30	5.66	32.08	0.9950	0.0667
CogX-Interp.*	15.59	28.60	31.88	0.9830	0.3333
CogVideoX	66.89	56.02	32.19	0.9738	0.7056
Latent Mix	45.17	60.02	31.99	0.9749	**0.7273**
Ours	**25.43**	**40.78**	**32.24**	**0.9850**	0.4722

Table 2. User Study Results.

	Temporal Consistency↑	Visual Quality↑	Video Dynamic↑
CogVideoX	3.34	3.62	3.68
Svd-Interp.	1.63	1.71	1.53
CogX-Interp.	2.22	2.08	2.17
Latent Mix	3.52	3.44	3.52
Ours	**4.30**	**4.15**	**4.10**

method with the state-of-the-art generative frame interpolation methods introduced in the baseline section. As shown in Fig. 5, the baseline interpolation methods may produce still results or generate content that is far away from the start frame and the end frame. The latent mix method blending the initial and final latent may result in artifacts in the end frame. Differently, the proposed method can generate the same start and end frames without noticeable differences. Due to the page limitation, we give more examples in Fig. 10 and the supplementary video.

As for the numerical comparison, as shown in Tab. 1, the proposed method shows a better visual quality and text-video alignment than previous methods. Besides, we also achieve a relatively higher score with both motion smoothness, video dynamic, and the MSE between the first frame and the last frame, which shows the advantage of the

proposed methods. We argue that although the latent mix method gives much dynamic video, the generated content might not be a looping one according to the MSE between the first and the last frame. Evaluating the looping videos using current automatic evaluation metrics is also difficult, so we conduct a subjective user study to prove the proposed method's effectiveness further. In detail, we invite 23 participants to rank ten questions across three aspects, totaling 690 opinions under five different methods. Each participant will be asked to rank the overall visual quality of the video, the consistency of the video frames, and how dynamic the video is, on a scale of 5 to 1. Finally, we calculate the average score of these opinions. As shown in Table 2, our method outperforms others in visual quality, temporal quality, and video dynamic.
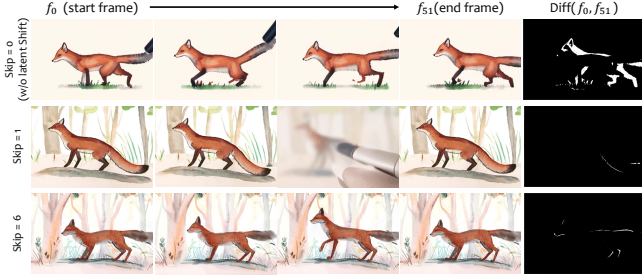
Fig. 6. **Ablation study on different latent skip.** The shift step in each denoising iteration will also influence the generated content.
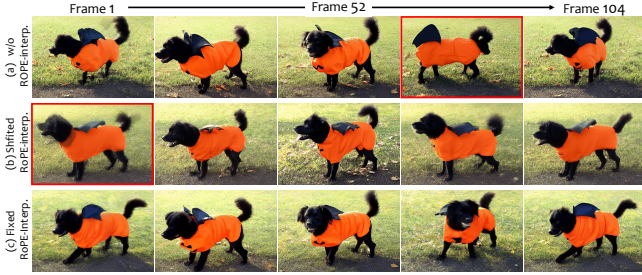

Fig. 7. **Ablation study on RoPE-Interp.** Under the implementation of latent shifting, different RoPE strategies can have a significant impact on the content of video generation.

4.3 Ablation Studies

We have given the example in Fig. 3 to validate the effectiveness of the proposed frame-invariance latent decoding. Here, we give more ablation studies on the ROPE-Interp. and the skip step in our latent shifting. When performing a latent shift, we can shift the latent s step for denoising, where a small step will be similar to the original inference. As shown in Fig. 6, when shifting the latent 6 steps in each denoising, the generated content is in a balance of the generated content and the motion. Differently, a small skip will show obversely artifacts.

We also conduct experiments on the RoPE interpolation. In the method, we give two different ways to utilize the interpolated RoPE. As shown in Fig. 7, the fixed RoPE-Interp performs well in our longer video looping generation, allowing each frame to be treated as the first frame during video generation, thereby achieving better looping results.

4.4 Applications on Longer Video Generation

Since the proposed latent shifting can naturally work for longer video generation, we also compare our method on longer video generation, where these baselines have been introduced in Section 2.3 for details. As shown in Figure 5, directly increasing the size of the latent causes video quality collapse. *Gen-L-Video* [Wang et al. 2023a] produces overly smooth transitions in the background and excessive changes in the direction of the seagull. *FreeNoise* [Qiu et al. 2023] tends to keep the seagull's orientation constant, the static nature of the image caused by latent shuffling is immediately apparent, and the phenomenon of the seagull having three legs also occurs.

Table 3. Comparing with other longer video generation methods.

	FVD↓	CLIP Score↑	Motion Smooth↑
Gen-L-Video [Wang et al. 2023a]	38.15	29.57	**98.86%**
FreeNoise [Qiu et al. 2023]	33.56	<u>32.34</u>	97.48%
FIFO [Kim et al. 2024]	41.25	32.15	96.83%
DiTCtrl [Cai et al. 2024]	<u>31.64</u>	32.13	97.89%
Ours	**29.89**	**32.43**	<u>98.04%</u>

Although *FIFO* [Kim et al. 2024] achieves better motion changes and video coherence, the issues of the seagull changing direction twice in a row and having three legs persist. *DiTCtrl* [Cai et al. 2024] improves the seagull's orientation issue, but still has problems with the defective generation of the seagull's head in the first frame and the three-legged issue. In contrast, the proposed method maintains the seagull's orientation while ensuring coherent video motion. It does not exhibit the issue of the seagull having three legs, thereby achieving superior long video generation. We give the full comparison in the supplementary video. As for the numerical comparison, we conduct the experiments on the same prompts of our looping video generation and calculate the main numerical results in Tab. 3 utilizing the well-known metrics from previous studies [Cai et al. 2024; Qiu et al. 2023].

4.5 Limitations

Since our method is a training-free method based on the pre-trained video diffusion model, our motion prior might be influenced by the pre-trained video diffusion model. As shown in Fig. 8, we give the results of the successive frame of the generated illustration video. However, the generated dress might not be consistent in the generated results and does not show obvious movement. We argue that this is because of the issues of the motion prior in the pre-trained video diffusion model we use. A better latent diffusion model [Brooks et al. 2024; Team 2024b] might work better.

Fig. 8. **Limitation**. The generated results might not show a very smooth video in the customized domain, *e.g.*, the illustration, restricted by the pre-trained text-to-video diffusion model.

5 CONCLUSION

We represent a novel and innovative approach to generating seamlessly looping videos directly from text descriptions without the need for user annotations. This is achieved by repurposing a pre-trained text-to-video latent diffusion model with inference latent modification. In detail, considering each frame should be considered

equally in the looping video, we construct a latent cycle and design latent shift to utilize the abilities of the video diffusion model's multi-frame latent denoising in each step, which further expands the scope of seamless looping video generation beyond the limitations of the video diffusion model's context. Besides, we introduce the frame-invariant latent decoding and RoPE-interpolation to further increase the performance. Compared to previous cinemagraphs, Mobius has a distinct advantage as it does not rely on an image for appearance, thus allowing for more dynamic motion and enhanced visual quality in the generated videos. Through multiple experiments and comparisons, the effectiveness of this method has been verified across different scenarios even on the application of longer video generation task.

REFERENCES

Aseem Agarwala, Ke Colin Zheng, Chris Pal, Maneesh Agrawala, Michael Cohen, Brian Curless, David Salesin, and Richard Szeliski. 2005. Panoramic video textures. In *ACM SIGGRAPH 2005 Papers*. 821–827.

Jiamin Bai, Aseem Agarwala, Maneesh Agrawala, and Ravi Ramamoorthi. 2013. Automatic cinemagraph portraits. In *Proceedings of the Eurographics Symposium on Rendering* (Zaragoza, Spain) *(EGSR '13)*. Eurographics Association, Goslar, DEU, 17–25. https://doi.org/10.1111/cgf.12147

Hugo Bertiche, Niloy J Mitra, Kuldeep Kulkarni, Chun-Hao P Huang, Tuanfeng Y Wang, Meysam Madadi, Sergio Escalera, and Duygu Ceylan. 2023. Blowing in the wind: Cyclenet for human cinemagraphs from still images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 459–468.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023).

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators. (2024). https://openai.com/research/video-generation-models-as-world-simulators

Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. 2024. DiTCtrl: Exploring Attention Control in Multi-Modal Diffusion Transformer for Tuning-Free Multi-Prompt Longer Video Generation. *arXiv preprint arXiv:2412.18597* (2024).

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22560–22570.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. 2023. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512* (2023).

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7310–7320.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725* (2023).

Tavi Halperin, Hanit Hakim, Orestis Vantzos, Gershon Hochman, Netai Benaim, Lior Sassy, Michael Kupchik, Ofir Bibi, and Ohad Fried. 2021. Endless loops: detecting and animating periodic patterns in still images. *ACM Transactions on graphics (TOG)* 40, 4 (2021), 1–12.

Mingming He, Jing Liao, Pedro V. Sander, and Hugues Hoppe. 2017. Gigapixel Panorama Video Loops. *ACM Trans. Graph.* 37, 1, Article 3 (Nov. 2017), 15 pages. https://doi.org/10.1145/3144455

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Aleksander Holynski, Brian L Curless, Steven M Seitz, and Richard Szeliski. 2021. Animating pictures with eulerian motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5810–5819.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. 2024. FIFO-Diffusion: Generating Infinite Videos from Text without Training. *arXiv preprint arXiv:2405.11473* (2024).

Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. 2024. Generative image dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24142–24153.

Zicheng Liao, Neel Joshi, and Hugues Hoppe. 2013. Automated video looping with progressive dynamism. *ACM Transactions on Graphics (TOG)* 32, 4 (2013), 1–10.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. GPT understands, too. *AI Open* (2023).

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22139–22149.

Aniruddha Mahapatra, Aliaksandr Siarohin, Hsin-Ying Lee, Sergey Tulyakov, and Jun-Yan Zhu. 2023. Synthesizing Artistic Cinemagraphs from Text. *arXiv preprint arXiv:2307.03190* (2023).

Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. 2024. MOFA-Video: Controllable Image Animation via Generative Motion Field Adaptions in Frozen Image-to-Video Diffusion Model. *arXiv preprint arXiv:2405.20222* (2024).

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4195–4205.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071* (2023).

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. *arXiv:2303.09535* (2023).

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. 2023. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169* (2023).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. 2024. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. 2024. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4733–4743.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.

Zhenxiong Tan, Xingyi Yang, Songhua Liu, and Xinchao Wang. 2024. Video-Infinity: Distributed Long Video Generation. *arXiv preprint arXiv:2406.16260* (2024).

Genmo Team. 2024a. Mochi 1. https://github.com/genmoai/models.

Hunyuan Video Team. 2024b. HunyuanVideo: A Systematic Framework For Large Video Generative Models. https://arxiv.org/abs/2412.03603

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018).

Fanyi Wang, Peng Liu, Haotian Hu, Dan Meng, Jingwen Su, Jinjin Xu, Yanhao Zhang, Xiaoming Ren, and Zhiwang Zhang. 2024a. LoopAnimate: Loopable Salient Object Animation. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*. 1–8.

Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. 2023a. Gen-l-video: Multi-text to long video generation via temporal co-denoising.

arXiv preprint arXiv:2305.18264 (2023).

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023b. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571* (2023).

Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. 2024b. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468* (2024).

Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher-Shlizerman, Aleksander Holynski, and Steven M Seitz. 2024c. Generative inbetweening: Adapting image-to-video models for keyframe interpolation. *arXiv preprint arXiv:2408.15239* (2024).

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. 2023. Freeinit: Bridging initialization gap in video diffusion models. *arXiv preprint arXiv:2312.07537* (2023).

Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. 2024. Tooncrafter: Generative cartoon interpolation. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 1–11.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. 2025. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*. Springer, 399–417.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).

Jiwen Yu, Xiaodong Cun, Chenyang Qi, Yong Zhang, Xintao Wang, Ying Shan, and Jian Zhang. 2023a. AnimateZero: Video Diffusion Models are Zero-Shot Image Animators. *arXiv preprint arXiv:2312.03793* (2023).

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. 2023b. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23174–23184.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2022. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. arXiv:2211.12194

"A seagull,…, walks along the shore, its sharp beak occasionally probing the sand for food. …"

Fig. 9. **Applications on Longer Video Generation**. We show some sampled frames here and the whole video is included in the supplementary video.
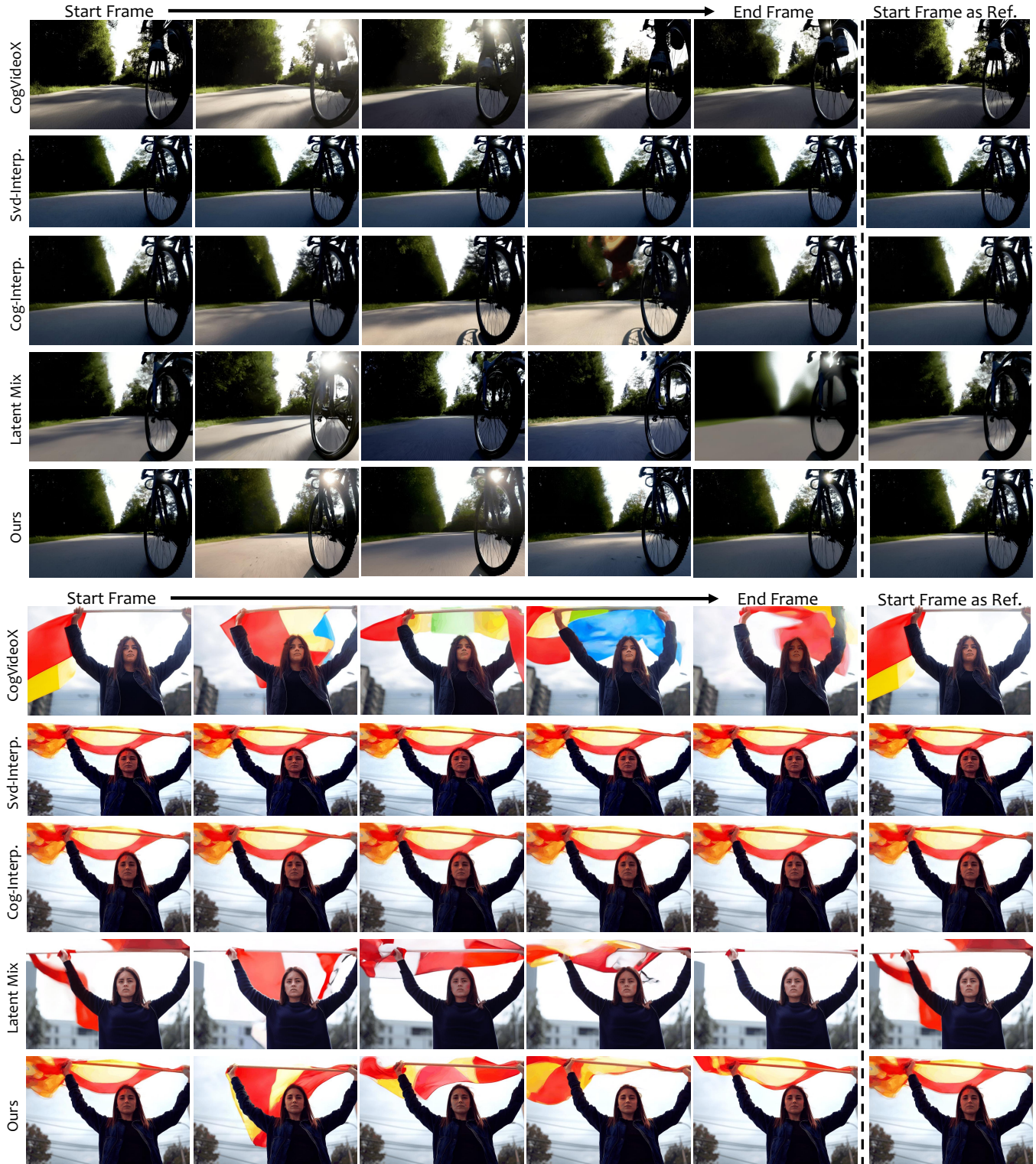
Fig. 10. More comparisons on the looping video generation.