ZoLA: Zero-Shot Creative Long Animation Generation with Short Video Model

Fu-Yun Wang[1], Zhaoyang Huang[2], Qiang Ma[4], Guanglu Song[4], Xudong Lu[1], Weikang Bian[1], Yijin Li[6], Yu Liu[3], and Hongsheng Li[1,3,5]

[1] MMLab CUHK
{fywang@link,hsli@ee}.cuhk.edu.hk
[2] Avolution AI
zhaoyanghuang@avolution.com
[3] Shanghai AI Lab
[4] SenseTime Research
[5] CPII under InnoHK
[6] Zhejiang University

Abstract. Although video generation has made great progress in capacity and controllability and is gaining increasing attention, currently available video generation models still make minimal progress in the video length they can generate. Due to the lack of well-annotated long video data, high training/inference cost, and flaws in the model designs, current video generation models can only generate videos of $2 \sim 4$ seconds, greatly limiting their applications and the creativity of users. We present ZoLA, a zero-shot method for creative long animation generation with short video diffusion models and even with short video consistency models (a new family of generative models known for the fast generation with high quality). In addition to the extension for long animation generation (dozens of seconds), ZoLA as a zero-shot method, can be easily combined with existing community adapters (developed only for image or short video models) for more innovative generation results, including control-guided animation generation/editing, motion customization/alternation, and multi-prompt conditioned animation generation, *etc*. And, importantly, all of these can be done with commonly affordable GPU (12 GB for 32-second animations) and inference time (90 seconds for denoising 32-second animations with consistency models). Experiments validate the effectiveness of ZoLA, bringing great potential for creative long animation generation. More details are available at https://gen-l-2.github.io/.

1 Introduction

Sora [7], released by OpenAI, achieves video generation of 20 to 60 seconds with ultra-high visual quality and spatiotemporal consistency. It shows video generation capabilities that far exceed all previous open source models, and has greatly

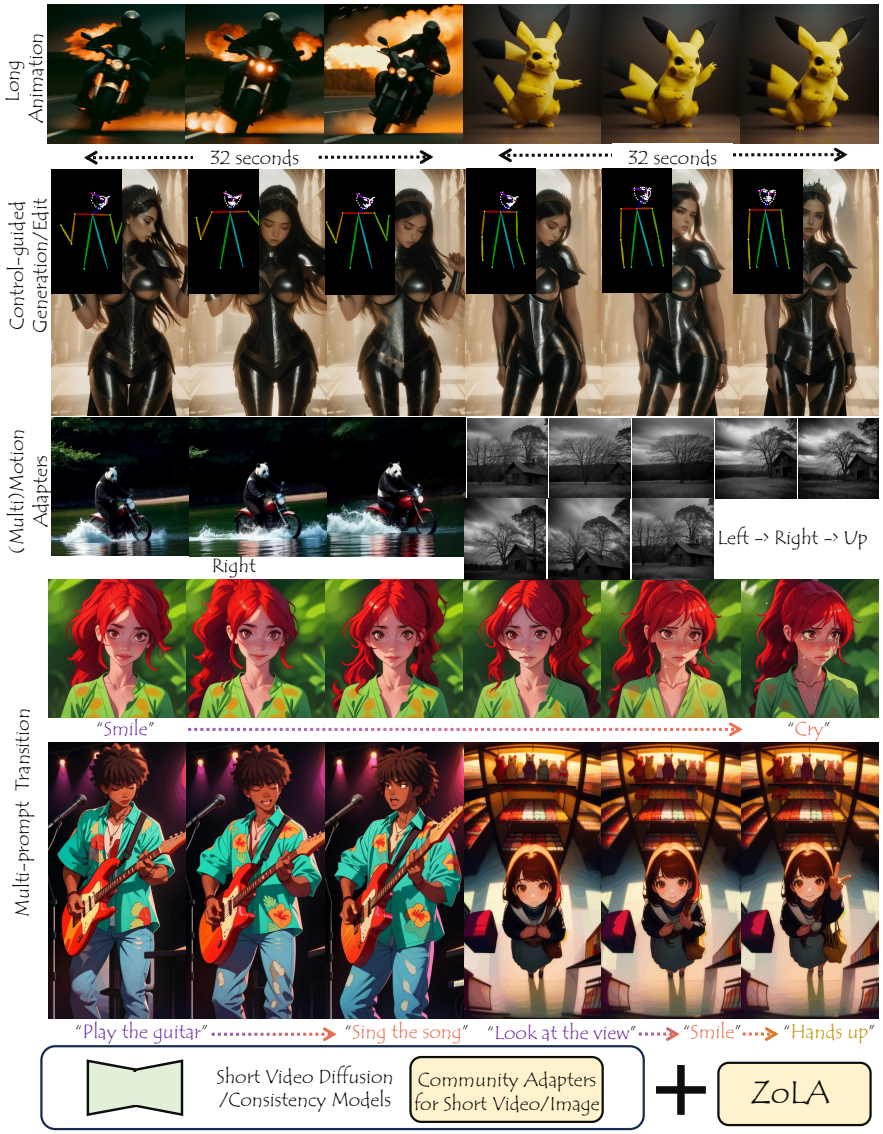

Fig. 1: ZoLA is a versatile zero-shot method for creative long animation generation. It can be combined with common video diffusion and even consistency models to achieve long animation generation, control-guided generation, motion customization/alternation, and multi-prompt conditioned animation generation with commonly affordable GPU memory and inference time.

attracted attention and discussion. Video generation models are considered to have the potential to become world simulators for understanding the physical laws of the world, and have become a crucial area of generative model research.

In contrast, the currently available mainstream video models (*e.g.*, Stable Video Diffusion) [5,6,9,10,16,24,25,31,34,41], although their visual quality and controllability are also improving, can usually only generate videos of 2 to 4 seconds, which limits their applications and the creativity of users. There are three main challenges that limit their generation length: 1) Low-quality text annotation of training data. Some video data only use the subtitles of the video as text annotations, which fail to reflect the real content of the video. High-quality annotation has been widely proven to be necessary to achieve high-fidelity generative models [1, 3, 5, 39]. 2) Training is expensive. For example, even though models with more parameters and larger capacity are constantly proposed, the current mainstream video models are usually developed based on Latent Diffusion Models (a.k.a, Stable Diffusion) [23] with 0.9B parameters, and decouple the space and time dimensional interactions to reduce the amount of computation [6, 25]. Even with such efforts, a GPU with 80GB memory can only allow one 4-second 576×320p video to be trained (with necessary optimization, *e.g.*, flash attention, mixed-precision). 3) Unreasonable model designs. Current video generation models usually only regard the video as a stack of image frames and neglect the temporal redundancy. This not only wastes the amount of computation but also leads to a large signal-to-noise ratio [5, 8], resulting in expensive training and inference costs and training difficulty.

These difficulties of long video generation show that it is barely affordable for most researchers or users (local usage), and therefore we propose an interesting zero-shot long animation generation scheme, termed as ZoLA (**Z**ero-Sh**O**t Creative **L**ong **A**nimation), which can be combined with current short video generation models (including diffusion models and consistency models) without training, and achieve creative long animations with commonly affordable inference time and computation resources. The core of our method lies in the length expansion at two dimensions: the spatiotemporal attention dimension and the denoising path dimension, achieving a good balance between the inference time and GPU memory cost. Besides, we show that the initialization of noise is important for the balance of freedom and stability of generation. In addition, we propose a noise travel strategy, which can effectively alleviate the degradation of visual quality and consistency caused by denoising conflicts. More details will be explicitly illustrated in the method section.

As shown in Fig. 1, our advantages include: 1) ZoLA can generate animations that are much longer but still maintain good temporal consistency and visual quality. 2) ZoLA supports more creative animation generation. 2.a) By combining editing algorithms (*e.g.*, SDEdit) [12,20] or layout conditions (optional) [21], we can achieve controllable video generation/editing. 2.b) Besides, ZoLA can combine the customized action adapters [18,38,46] of short videos to realize the motion customization and alternations of several motions in a longer animation. 2.c) We also achieve alternating control of multiple prompts. For example, controlling changes in a person's facial expression through changes in text prompts along the time. 3) ZoLA is efficient and commonly affordable. Compared with the direct generation method, our method has a significant reduction in GPU

memory. For example, using the same mainstream video model structure, directly generating a 32-second 512×512p video usually requires more than 40GB of GPU memory. ZoLA only requires 12GB of GPU memory. Besides, ZoLA can be well integrated with various sampling acceleration methods. For example, using the video consistency model [34], we can generate videos with only 4 steps of sampling. We only need 26 seconds to generate an 8-second animation. In contrast, a video diffusion model with the same architecture usually takes 25 seconds to generate a 2-second video (tested on A800).

Although, as a zero-shot generation strategy, the generation quality of ZoLA is upper-bounded by the base short video models and thus cannot achieve the same effect as long videos like sora. But on one hand, we hope that this idea can support creative and longer animation generation to a certain extent in the era with only short video models available for most. On the other hand, we consider our method to be relatively versatile and capable of integrating with different model architectures. This versatility ensures that, even as more powerful models emerge, our strategy can be applied to harness these advancements for more creative outcomes. We conducted extensive experimental comparisons with previous methods. Quantitative experimental metrics and user studies show that our method achieves better consistency and visual quality.

2 Related Works

Diffusion models. Diffusion models (a.k.a., score-based models) [14, 15, 19, 26, 27] have received increasing attention due to their amazing ability to generate highly detailed images. Currently, video diffusion models [6, 14, 24, 25, 32, 34, 35] generally extend image diffusion models by inserting temporal layers. These models are either trained by joint image-video tuning [16, 25] or by spatial weight freezing [6] to mitigate the poor annotations and visual quality of raw video data.
Long video generation. Previous works [9–11, 30, 32, 42] for long video generation commonly follow the auto-regressive mechanism with a temporal mask modeling technique. NUWA-XL proposes a hierechical way for long video generation [40]. However, little work validates their success on open-domain text-to-video generation. Additionally, the auto-regressive mechanism behind them has inevitable drawbacks: 1) Huge retraining cost. Due to the introduction of temporal mask modeling, they generally have to retrain the model to accept additional conditions. The retraining leads to additional training costs and can potentially cause the generation quality degradation. 2) Training-inference gap and flaws accumulation. They are trained to predict the next short video clip utilizing the real former short video clip as conditions, but they have to use short video clips generated by themselves to predict the next video clip. Flaws caused by the training-inference gap greatly accumulate in the auto-regressive process, leading to dramatic degradation. 3) Looping and inauthenticity. Due to the capacity, they predict the next video clip only with the information of the very last video clip they generated (*i.e.*, only one clip is fed as conditional inputs). This tends to cause looping and inauthenticity in their generated results.

3 Preliminaries

Diffusion models [15] perturb the data by gradually injecting noise to data $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$, which is formalized by a markov chain:

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \quad q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{\alpha_t}\boldsymbol{x}_{t-1}, \beta_t\boldsymbol{I}), \tag{1}$$

where β_t is the noise schedule and $\alpha_t = 1 - \beta_t$. The data can be generated by reversing this process, *i.e.* we gradually denoise to restore the original data. The diffusion model $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ parameterized by $\boldsymbol{\theta}$ is trained to approximate the reverse transition $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$, which is formulated as

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \frac{1}{\sqrt{\alpha_t}}\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\sqrt{\alpha_t}}\boldsymbol{\epsilon}, \tilde{\beta}_t\boldsymbol{I}) \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, and $\boldsymbol{\epsilon}$ is the noise injected to $\boldsymbol{x}_0$.

4 Methodology

4.1 Model Architectures

As we mentioned before, current video models are typically built upon the extension of pretrained image models (*e.g.*, Stable Diffusion). Most current open-source video generation models, follow the idea of adding temporal convolutions and temporal attention modules to modeling the temporal interactions. The features $\boldsymbol{z} \in \mathbb{R}^{b\times f\times h\times w\times c}$ are first rearranged to move all spatial dimensions into the batch dimension $\boldsymbol{z}' \in \mathbb{R}^{(b\times h\times w)\times f\times c}$ and then features at the same spatial position but different frames will be processed by temporal convolution or attention blocks. The newly added modules are then trained on video data or joint video-image training to align the denoising path of different frames. This design has been widely verified to ease the burden of training and achieves acceptable results [6, 25, 36, 43]. While this tricky design may also limit their performance upper-bound.

4.2 Problem Formulation

Our goal is to generate creative long animations at a commonly affordable cost, using only short video models. It is important to note that current video models generally regard videos as mere stacks of video frames, so we define the length of a video as the number of its frames. In ZoLA, we expand the length by extending in both the spatiotemporal attention dimension and the denoising path dimension. We assume that the given short video model is trained on videos of base length f (*i.e.*, the model has the capability to denoise videos of base length f). Since videos can be seen as a stack of video frames, then the model should have the ability to denoise any set of f images with temporal order and relationships.

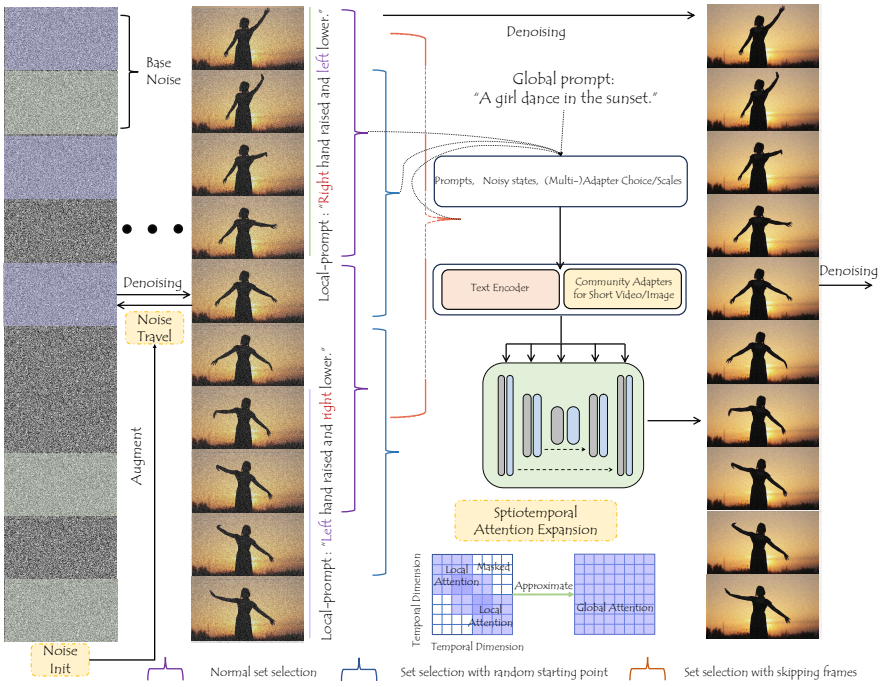

Fig. 2: Workflow of ZoLA. ZoLA extend the video generation length by expansion at two dimensions: spatiotemporal attention expansion and denoising path expansion. Special designed noise init strategy and noise init augmented noise travel strategy are additionally proposed to enhance temporal consistency and visual quality. In the above figure, we assume the base generation length $f = 2$ and the expanded length $f' = 4$, $f'' = 10$. Stride s for set selection is set to 2. Maximum frame skipping R is set to 2.

On this basis, we first adopt an expansion of the spatiotemporal attention dimension. Specifically, this is a kind of attention approximation operation that uses local attention interactions to approximate global temporal attention interactions. Through this method, the temporal interaction module still only needs to process interactions among the f features as the length they are trained with, thus minimizing the disruption to the model's generative capabilities. We assume that, based on this, the model is capable of denoising videos of length f'. We generally find that it works well when f' is not significantly larger than f. In an abstract sense, we have evolved a model that can only denoise videos of length f into a model that can denoise videos of length f' tolerating a certain of approximation errors.

Building on this, we continue with the expansion of the denoising path dimension. Note that we just mentioned videos can be seen as a series of image frames with temporal relationships. Assuming the video length we want to denoise is f'', this is equivalent to a set of f'' image frames. Therefore, we just need to continuously extract sets of f' sequentially related images from the f'' images

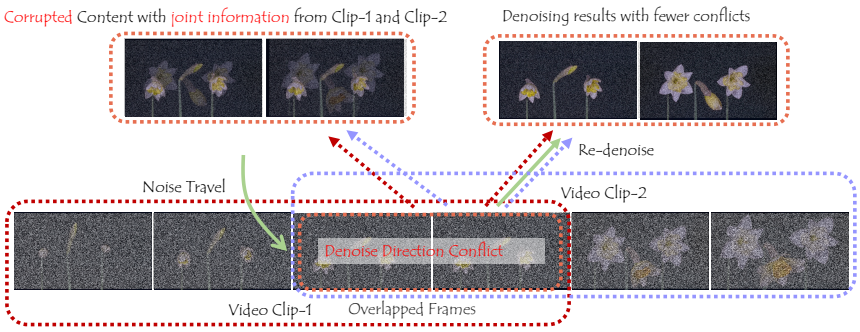

Fig. 3: Noise travel. Noise travel allows the model to reverse the denoising step after merging the conflicted denoising directions and re-denoise again. This process enhances the information transfer between different clips and alleviates potential denoising conflicts, thus further promoting global consistency and generation quality. In the figure, each set has 4 frames (*i.e.*, $f' = 4$) and stride for set selection is 2 (*i.e.*, $s = 2$). Therefore, the adjacent sets have two frames overlapped.

for denoising until all image frames have been denoised. Hence, the problem becomes the Set Cover Problem (SCP).

In summary, ZoLA is built upon two length expansions: spatiotemporal attention dimensional expansion ($f \to f'$) and denoising path dimensional expansion ($f' \to f''$), where $f'' > f' > f$. Now we have demonstrated the motivations and high-level ideas. In the following, we will go into the details of their implementations.

4.3 Length Expansion

Spatiotemporal attention expansion ($f \to f'$). Inspired by previous work [6], the core idea of spatiotemporal attention expansion is to extend the video length by approximating global attention interactions with local attention interactions. In other words, it can be perceived as applying spatiotemporal attention convolutionally at the temporal dimension. To be specific, before attending the spatiotemporal attention, given the feature $\boldsymbol{z} \in \mathbb{R}^{b \times f' \times h \times w \times c}$, the features are first split into windows with window size f equal to the base length with sliding stride s ($s \leq f$). Hence, the feature $\boldsymbol{z}$ is transformed into a set of features $\boldsymbol{z}^i \in \mathbb{R}^{b \times f \times h \times w \times c}$ with base length f, namely

$$\boldsymbol{z}^i = \boldsymbol{z}[:, i \times s : i \times s + f], \quad i = 0, 1, \dots \tag{3}$$

where $\boldsymbol{z}[:, i \times s : i \times s + f]$ is a numpy style notation, representing the slicing of $\boldsymbol{z}$ from frame $i \times s$ to frame $i \times s + f$. Then the split sets of features are sent to normal spatiotemporal modules for independent processing. After that, the outputs of these sets of features are merged back into the original shape through weighted interpolation. We set $s = \lfloor f/2 \rfloor$. To better preserve the consistency and alleviate the possible sudden changes at the sides of windows, we set the

merge weights inside the window decays as the deviation from the center frame. That is,

$$\boldsymbol{w}^i[j] = \exp\left(-\frac{1}{2}\left|j - \lfloor f/2 \rfloor\right|\right), \tag{4}$$

where $j = 0, 1, 2, \ldots, f-1$ is the frame index inside the window.

Denoising path expansion ($f' \to f''$). As we discussed before, a video can be perceived as a set of sequential images, and therefore after the spatiotemporal attention expansion, we now have a tool to denoise any sets of f' images with temporal relations. To denoising longer videos (f''), we just need to solve the Set Cover Problem that we constantly choose subsets of length f' until all f'' images are covered. A normal set selection strategy is following the sliding window. That is, from 0 to f'', we constantly select the set with length f' in a sliding way with stride s ($s \leq f'$). Namely,

$$\boldsymbol{v}^i = \boldsymbol{v}[:, i \times s : i \times s + f'], \quad i = 0, 1, \ldots . \tag{5}$$

However, this kind of set selection strategy confines the information interactions within the sets. This will cause sudden changes and inconsistency in the corners of sets. For instance, Gen-L [33] applies a similar idea termed temporal co-denoising while they have to apply small strides to have more stable results.

Note that the denoising process is an iterative process, with multiple steps for gradually reconstructing the target video from white noise. Therefore, we instead of applying the fixed selection of sets, propose to apply different set selections at different denoising timesteps. To be specific, this can be achieved through randomly sampling the starting frame index k for set selection. In this way, the i-th set selected for denoising can be represented as

$$\boldsymbol{v}^i = \boldsymbol{v}[:, k + s \times i : k + s \times i + f'], \quad k \sim \text{Uniform}\{0, 1, 2, \ldots f' - 1\}. \tag{6}$$

Potential indices exceeding f'' will be left shifted to satisfy the length condition. In this way, frames will interact in a more flexible way, and we can apply larger strides to reduce the inference cost (since the number of sets is $\lceil \frac{f''-f'}{s} \rceil + 1$).

To allow for more interactions of long-range frames and thus enhance long-term consistency, we additionally propose to allow for set selections with skipping frames. That is, to select an ordered subset from a large set, we could sequentially sample one element (image) by skipping several elements. This is reasonable if we consider the concept of frames-per-second (fps) in videos. Larger fps cause the video to have a more dense frame distribution and lower fps cause the video to have a more sparse frame distribution. However, as long as the fps belongs to reasonable intervals, it will not affect the viewing experience. To be specific, the i-th set selected for denoising can be represented as,

$$\begin{aligned} \boldsymbol{v}^i &= \boldsymbol{v}[:, k + s \times i : k + s \times i + f' \times r : r], \quad i = 0, 1, 2, \ldots \\ k &\sim \text{Uniform}\{0, 1, \ldots, f' - 1\}, \quad r \sim \text{Uniform}\{1, \ldots, R\}, \end{aligned} \tag{7}$$

where $\boldsymbol{v}[:, k + s \times i : k + s \times i + f' \times r : r]$ is a numpy style notation of slicing with stride r. $R \in \mathbb{N}^+$ is the maximum number of frames allowing for skipping. The

potential indices exceeding f'' will also be left shifted to satisfy the length limit. We conduct this set selection strategy to denoise selected sets independently until all frames are denoised at a specific timestep. The denoising results for those frames being selected more than once will be averaged as the actual results. Following the proof in previous work [2, 33], as long as the denoising results are close enough, we can approximate the longer video denoising path with short video denoising ability.

4.4 Noise Design

Noise init. The initial noise for generation contains crucial information for generating videos, as it determines the basic structures and motions in videos. The spatiotemporal modules in the video generation models are trained to align a fixed number of frames f. Therefore, when extended to longer videos with length f'', we empirically find that they may fail to align the much longer noises due to too much flexibility in higher dimensional space even above proposed length expansion strategies are employed. Therefore, it is essential to reduce the flexibility of the initial noise. A vanilla way is to repeatedly concatenate the noise with length f along the time axis until it reaches f''. In this way, the initial noises, though constructed as a high-dimensional white noise, are actually confined in a low-dimensional manifold. However, we find that this typically causes the model to generate repeated motions. To this end, we propose to simultaneously sample a random noise with extended length f'' and a base noise with the base length f. We denote them as $\boldsymbol{\epsilon}'' \in \mathbb{R}^{f'' \times h \times w \times c_{\text{latent}}}$ and $\boldsymbol{\epsilon} \in \mathbb{R}^{f \times h \times \times c_{\text{latent}}}$. The random noise is randomly replaced by the base noise in the frame level. Specifically, we repeatedly concatenate the base noise at the frame level to extend it to the length of f'' and add it to the random nose $\boldsymbol{\epsilon}''$ with mask $\boldsymbol{m} \in \mathbb{R}^{f \times 1 \times 1 \times 1}$. Then, the noise is initialized as

$$\boldsymbol{v}_T = (\mathbf{1} - \boldsymbol{m}) \odot \boldsymbol{\epsilon}'' + \boldsymbol{m} \odot \text{repeat}(\boldsymbol{\epsilon}), \tag{8}$$

In this way, the mask ratio controls the generation flexibility and stability. Larger mask ratios allow for more generation stability, but too large mask ratios cause the motion corrupted, which we will show in the ablation study (Fig. 5).

Noise travel (augmented with noise init). As we mentioned following Eq. 7, the denoising path expansion achieves good results if and only if the denoising directions of overlapped frames in different sets are close enough. However, this is usually not the case. Let us consider a random initialized noise where the sets selected have no information about each other at the beginning of denoising. Therefore, it is typical for them to have denoising results conflicts. Although many conflicts could be alleviated at the later denoising steps, heavy conflicts could lead to the noisy state wrongly entering into the low-density region of score manifold [27], which would further corrupt the denoising process, causing generation failure or degraded generation quality. As shown in Fig. 3, due to the incomplete perception of the whole video, the denoising directions of the two different sets for the same frames (frames in the middle) are actually pointing to

two different results, thus corrupting the denoising results. The corrupted results would even further corrupt the whole denoising process. However, it should be noted that although corrupted, the overlapped frames now contain information from both sets as illustrated in Fig. 3. To this end, we propose to reverse the denoising step back and hence provide a second chance for the model to denoise the noisy state at the same timestep. In this way, the denoising directions of both sets are now influenced by each other, making them more likely to achieve compatible denoising results. Assume we have obtained the denoising result $\boldsymbol{v}_{t-1}$ from $\boldsymbol{v}_t$. To implement the noise travel, we sample a noise $\boldsymbol{\epsilon}''$ and add it with $\boldsymbol{v}_{t-1}$. We also find that applying our proposed noise init strategy for sampling the noise ϵ at later denoising steps can further enhance the generation color consistency. Besides, in practice, we would typically let the denoising continue several steps (denoted as jump length L) instead of one step for more global information integration before noise travel, namely

$$\widehat{\boldsymbol{v}}_t = \sqrt{\prod_{i=t-L+1}^{t} \alpha_i}\boldsymbol{v}_{t-L} + \sqrt{1-\prod_{i=t-L+1}^{t} \alpha_i}\left[(\mathbf{1}-\boldsymbol{m})\odot\boldsymbol{\epsilon}'' + \boldsymbol{m}\odot\text{repeat}(\boldsymbol{\epsilon})\right]. \tag{9}$$

We only apply the noise travel at early denoising steps since they are more likely to cause denoising conflicts. We show that the nature of noise travel is equivalent to the score distillation sampling (SDS) [22] widely applied in 3D generation (supplementary).

4.5 Application: Creative Long Animation

Control-guided generation/edit. For achieving control-guided video generation, we norm the extracted control sequences (*e.g.*, pose sequences) and then feed them into the ControlNet [44] trained on images. For editing, we follow the SDEdit [20], perturbing the input video with white noise into certain noisy timesteps and then denoise it with the new prompt. We find that extracting control sequences and then applying ControlNet with SDEdit can better preserve the layout of the source video. For the noise added for editing, we find that it also works to apply the noise init strategy instead of pure white noise.

Motion customization/alternation. Motion customization [18, 46] has been vastly tested on short video generations. That is, several short videos with similar motions are firstly collected to train plug-and-play adapters (*e.g.*, LoRA [17]). Then, the model can generate videos with similar motions. When the user provides motion intervals (*e.g.*, which motion to use at which time interval), ZoLA achieve motion customization/alternation through pre-loading the adapters into memory and conditionally adjusting the insertion weights depending on the set of frames selected for denoising. Take the camera pan as an example, when the selected set of frames contains half the number of frames tagged with 'left' and half frames tagged with 'up', then insertion weights for the 'left' adapter and 'up' adapter are all $\frac{1}{2}\alpha$, where α is the hyper-parameter scale factor.

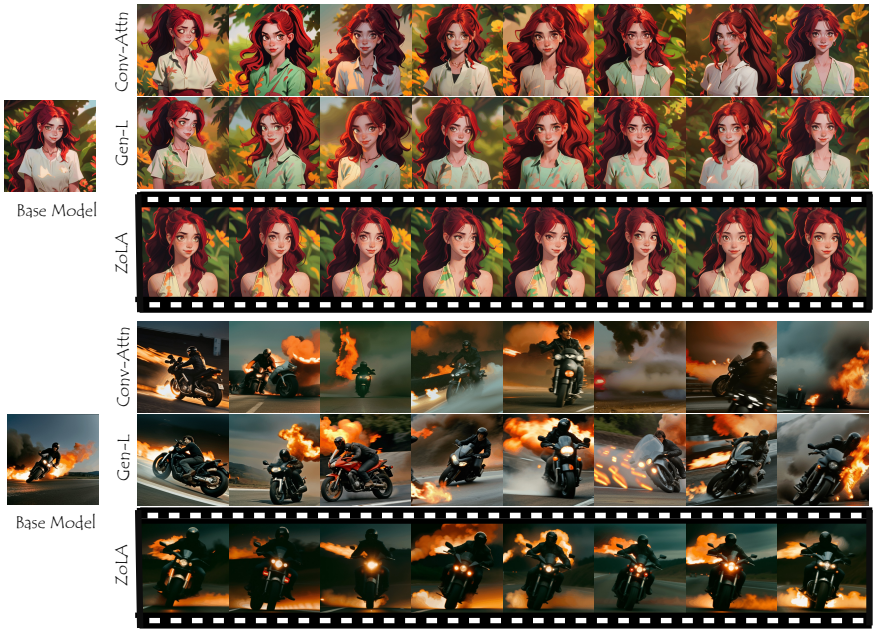

Fig. 4: Qualitative comparison. All videos are 256 frames (about 32 seconds). To better presentation in the paper, we sample one frame every 32 frames. Our method achieves much better long-term consistency compared to other methods, which suffer from background sudden changes or even foreground identity changes.

Multi-prompt transition. Except for the global prompt to determine the main content of the long animation, ZoLA can also accept a sets of local-prompt and their corresponding intervals. Note that previous work [33] also proposes multi-prompt transition through text embedding interpolations, we find that interpolation will potentially corrupt the text embedding. For a selected set of frames, we simply choose the interval containing most same frames with the selected set, and apply its corresponding prompt as the prompt for denoising.

5 Experiments

5.1 Experimental Setup

Protocols. For quantitative experiments, We selected ten foundational models from Civitai [7]. For each, we collect 10 prompts for generation. To stabilize the metric computation, we generate 10 videos for each prompt. This is reasonable. Considering that in the class-conditioned generation, we typically generate dozens of samples for each class. Thus, we generate 1,000 videos in total. Each

[7] https://civitai.com/

Table 1: Quantitative comparison of long video generation method. ↑ indicates "higher is better" and ↓ indicates "lower is better".

Method	Metrics					User study	
	CLIP-SIM ↑	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓	Consistency ↑	Visual Quality↑
Short	0.9693	0.6255	17.79	0.2499	/	/	/
Conv-Attn [6]	0.8895	0.4330	12.64	0.5892	526.1	10.2%	20.9%
Gen-L [33]	0.8891	0.4391	12.58	0.5897	440.3	21.5%	30.2%
ZoLA	**0.9443**	**0.6091**	**15.44**	**0.3374**	**387.1**	**68.3%**	**48.9%**

video contains 256 frames (32 seconds). The detailed models and prompts are listed in the supplementary.

Metrics. We apply FVD [29] to measure the distribution distance between the long animations and the short animations. We apply SSIM [37], CLIP-SIM [13], and LPIPS [45] to measure the long animation consistency in three levels: perceived quality, semantic similarity, and perceptual similarity. We also conduct a user study to measure the animation consistency and visual quality. For FVD, we uniformly sample 16 frames from long animations to match the short animations with the length of 16. For SSIM, LPIPS, and CLIP-SIM, we evaluated the longer videos for quality, perceptual detail, and semantic similarity by randomly comparing adjacent frames. Details of the user study are listed in the supplementary.

5.2 Qualitative Comparison.

We present a qualitative comparison in Fig. 4, which includes a relatively static close-up video of a person and a relatively dynamic motorcycle racing scene. Both videos consist of 256 frames, and we extract one frame every 32 frames to better showcase the results. The Conv-Attn captures short-range visual consistency in scenes, but it is evident that there are many inconsistencies in the background and even in the identity and movements of the foreground characters. The Gen-L better captures information from adjacent segments and maintains the consistency of the main character in relatively distant frames. However, we still observe many abrupt changes and long-distance inconsistencies. Our method achieved the best results, maintaining the best long-range consistency of the characters, movements, and backgrounds.

5.3 Quantitative Results.

Table 1 illustrates the quantitative experimental results. ZoLA demonstrated superior performance, closely approaching the upper bound set by Short (which denotes the original generation ability of the short video model for short videos), with a CLIP-SIM score of 0.9443, SSIM of 0.6091, and PSNR of 15.44. ZoLA also achieved the best in LPIPS (0.3374) and FVD (387.1), indicating higher visual fidelity and temporal consistency. User study further supports these findings, with ZoLA leading in Consistency (68.3%) [4, 28] and Visual Quality (48.9%).

Table 2: Quantitative ablation study onthe key components of ZoLA.

	CLIPSIM ↑	LPIPS ↓	Consistency ↑	Motion Quality ↑
TCD	0.880	0.533	14.2%	22.2%
LE	0.912	0.423	17.8%	23.8%
LE+NI	0.961	0.212	33.2%	18.4%
LE+NI+NT	0.953	0.249	34.8%	35.6%

Table 3: Computation complexity comparison.

	VRAM (GB)		Time (Sec)	
Video Length (Sec)	8	32	8	32
Conv-Attn	11.7	45.4	189.6	712.6
Gen-L	**5.1**	**5.1**	312.9	1284.4
ZoLA	7.3	11.8	275.2	934.5
ZoLA-CM	7.3	11.8	**26.1**	**82.3**

Fig. 5: Visual examples of ablation study on noise init.

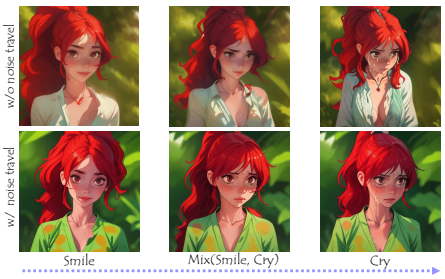

Fig. 6: Visual examples of ablation study on noise travel.

These results underscore the effectiveness of our method in generating high-quality long animation.

5.4 Ablation Study.

Effectiveness of each component. We conduct the quantitative ablation study with the same prompt set to generate 500 videos (5 videos for each) with 64 frames for each baseline. A user study with 500 comparison sets is conducted to evaluate the consistency and motion quality. Since the temporal co-denoising (TCD) from previous work [33] is an important baseline, we begin with it as the baseline. As shown in the table, the length expansion (LE) from ZoLA improves significantly against it in generation consistency. Additionally, ZoLA noise designs also contribute to the generation consistency and motion quality. Specifically, noise init (NI) contributes most to the consistency but can potentially degrade the motion quality. Noise travel (NT) works to both enhance consistency and visual/motion quality. Note that although the LPIPS and CLIPSIM indicate that adding NT to NI the consistency is lower, but our user study and subjective observation show its positive impact for consistency and motion/visual quality.

Visual examples of ablation study on the importance of noise. Fig. 5 illustrates the results of our ablation study focusing on noise init. Videos are generated using mask ratios of 0.0, 0.5, and 1.0, respectively. The lower mask ratios correspond to increased frame-to-frame flexibility, whereas higher ratios result in greater frame stability. As depicted, a mask ratio of 0.5 yielded the most favorable outcome. The 0.0 ratio lacked sufficient consistency, and a ratio

Fig. 7: Long animation generation (8 seconds) results comparison of ZoLA with modelscope and the official API. The left is to show the consistency. The right is to show the visual quality.

of 1.0 led to noticeably constrained motion. Fig. 6 illustrates an ablation study on the impact of noise travel in a multi-prompt generation example. Without noise travel, the generated results showed color degradation and noticeable flaws, especially evident in the top right corner of the image. Conversely, employing noise travel significantly enhanced the video quality.

5.5 Discussion

Computation complexity. We report the inference time and peak GPU memory (VRAM) usage of different methods for generating 512×512p animations with 8 and 32 seconds. All experiments are tested on a single A800. We exclude the decoding process (reconstruct the frames from latents) and only compare the denoising process. The results are shown in Table. 3, it indicates that ZoLA achieves a great balance in inference time and GPU memory compared with other methods. The combination with consistency models makes it significantly more efficient for the generation with commonly affordable GPU usage.
Other models with ZoLA (Versatility of ZoLA). To show the versatility of ZoLA, we additionally test ZoLA on modelscope [36]. As shown in Fig, 7, results with ZoLA are significantly better than the official API. Dealing with little with the architecture, ZoLA can be combined with various video models. Even when stronger models are proposed in the future, we believe they can benefit from ZoLA for longer creative animation generation.

6 Conclusion

We propose a novel inference strategy ZoLA, achieving high-quality long video generation with video diffusion models only trained to generate short videos. We first introduce quadratic temporal expansion for extending the length of video generation. Then we introduce the temporal consistent noise scheduler for alleviating potential approximation errors caused by quadratic expansion and better global consistency. Qualitative comparisons and quantitative experiments validate the effectiveness of ZoLA.
Limitations: Due to the zero-shot nature of ZoLA and generation stochasticity, it is difficult to align very long animations. ZoLA might encounter inconsistency or abrupt changes in partial details.

Acknowledge:

This project is funded in part by National Key R&D Program of China Project 2022ZD0161100, by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, by Smart Traffic Fund PSRI/76/2311/PR, by RGC General Research Fund Project 14204021. Hongsheng Li is a PI of CPII under the InnoHK.

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV. pp. 1728–1738 (2021)
2. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
4. Bian, W., Huang, Z., Shi, X., Dong, Y., Li, Y., Li, H.: Context-pips: Persistent independent particles demands context features. Advances in Neural Information Processing Systems **36** (2024)
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
6. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR. pp. 22563–22575 (2023)
7. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/research/video-generation-models-as-world-simulators
8. Chen, T.: On the importance of noise scheduling for diffusion models. arXiv preprint arXiv:2301.10972 (2023)
9. Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.B., Parikh, D.: Long video generation with time-agnostic vqgan and time-sensitive transformer. arXiv preprint arXiv:2204.03638 (2022)
10. Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C., Wood, F.: Flexible diffusion modeling of long videos. arXiv preprint arXiv:2205.11495 (2022)
11. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221 (2022)
12. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
13. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
14. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)

15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
16. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
18. Jeong, H., Park, G.Y., Ye, J.C.: Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. arXiv preprint arXiv:2312.00845 (2023)
19. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. NeurIPS **35**, 26565–26577 (2022)
20. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
21. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
22. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
24. Shi, X., Huang, Z., Wang, F.Y., Bian, W., Li, D., Zhang, Y., Zhang, M., Cheung, K.C., See, S., Qin, H., et al.: Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
25. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
26. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
27. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
28. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: ECCV. pp. 402–419. Springer (2020)
29. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
30. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual description. arXiv preprint arXiv:2210.02399 (2022)
31. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Masked conditional video diffusion for prediction, generation, and interpolation. arXiv preprint arXiv:2205.09853 (2022)
32. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. NeurIPS (2022)
33. Wang, F.Y., Chen, W., Song, G., Ye, H.J., Liu, Y., Li, H.: Gen-l-video: Multi-text to long video generation via temporal co-denoising. arXiv preprint arXiv:2305.18264 (2023)

34. Wang, F.Y., Huang, Z., Shi, X., Bian, W., Song, G., Liu, Y., Li, H.: Animatelcm: Accelerating the animation of personalized diffusion models and adapters with decoupled consistency learning. arXiv preprint arXiv:2402.00769 (2024)
35. Wang, F.Y., Wu, X., Huang, Z., Shi, X., Shen, D., Song, G., Liu, Y., Li, H.: Be-your-outpainter: Mastering video outpainting through input-specific adaptation. arXiv preprint arXiv:2403.13745 (2024)
36. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
37. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
38. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
39. Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., Guo, B.: Advancing high-resolution video-language representation with large-scale video transcriptions. In: CVPR. pp. 5036–5045 (2022)
40. Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., et al.: Nuwa-xl: Diffusion over diffusion for extremely long video generation. arXiv preprint arXiv:2303.12346 (2023)
41. Yu, S., Sohn, K., Kim, S., Shin, J.: Video probabilistic diffusion models in projected latent space. arXiv preprint arXiv:2302.07685 (2023)
42. Zeng, Y., Wei, G., Zheng, J., Zou, J., Wei, Y., Zhang, Y., Li, H.: Make pixels dance: High-dynamic video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8850–8860 (2024)
43. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. arXiv preprint arXiv:2309.15818 (2023)
44. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
46. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. arXiv preprint arXiv:2310.08465 (2023)