Video Generation with Consistency Tuning

Chaoyi Wang, Yaozhe Song, Yafeng Zhang, Jun Pei, Lijie Xia, Jianpo Liu
Chinese Academy of Sciences, Shanghai Institute of Microsystem and Information Technology
chaoyiwang, YaozheSong, acampus, peijun, xialj, liujp@mail.sim.ac.cn

Abstract

Currently, various studies have been exploring generation of long videos. However, the generated frames in these videos often exhibit jitter and noise. Therefore, in order to generate the videos without these noise, we propose a novel framework composed of four modules: separate tuning module, average fusion module, combined tuning module, and inter-frame consistency module. By applying our newly proposed modules subsequently, the consistency of the background and foreground in each video frames is optimized. Besides, the experimental results demonstrate that videos generated by our method exhibit a high quality in comparison of the state-of-the-art methods [14, 1, 15].

1. Introduction

Recently, diffusion models have achieved great success in handling complex and large-scale image datasets [2], these methods have shown great potential to model video distribution much better with scalability both in terms of spatial resolution and temporal duration [6]. Besides, a current method extends off-the-shelf short video diffusion models for generating long video [14].

Despite advances in video generation, the generated videos still present certain shortcomings. For example, unnatural jumps always occur between frames of generated videos. Therefore, we propose an automated video processing framework to address the aforementioned shortcomings, which includes four modules: separate diffusion module, average fusion module, combined tuning module and inter-frame consistency module.

2. Related Work

In recent years, the generation of long videos based on text has drawn tremendous attention, leading to various attempts to tackle the challenges associated with this task [13, 16, 3]. Denoising Diffusion Probabilistic Model (DDPM) [7] and its variant Denoising Diffusion Implicit Model (DDIM) [12] have been widely used for text-to-image generation, including MCVD [13], FDM [4], LVDM [5], PVDM [16], Gen-L-Video [14] and so on. In the presence of high quality text conditioned image generation models, several recent works have focused on utilizing additional control signals for generation or editing existing images [14]. However, these methods tend to have some limitations for practical applications that maintaining consistency is challenging as frame count increases in the video generation.

Therefore, in order to generate videos with consistency between frames, We demonstrate the above four modules in our proposed framework.

3. Methods

We develop a framework to generate video with high quality context and smooth frames composed of four modules in sequence, separate tuning module, average fusion module, combined tuning module and inter-frame consistency module. The detailed implementation of the methods are shown in figure 1.

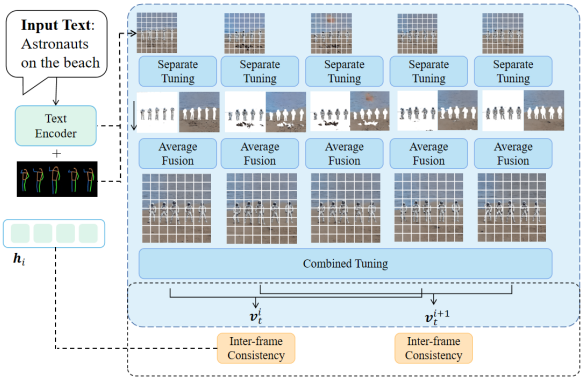

Figure 1. A high level overview of the proposed video generation approach.

3.1. Separate Tuning Module

Given the input prompts of the hidden embeddings $\boldsymbol{h}$, and the corresponding condition $\boldsymbol{c}$, image frames $\boldsymbol{r}$ are cre-

ated by

$$\boldsymbol{r} = \mathcal{F}_0(\boldsymbol{h}, \boldsymbol{c}) \tag{1}$$

where $\mathcal{F}_0$ denotes the function of the diffusion and denosing process.

In details, in the whole diffusion and denoising process, the diffusion models [11] are trained through latent variable representation of the form $p_\theta(\boldsymbol{x}) := \int p_\theta(\boldsymbol{x}_{0:T}) d_{\boldsymbol{x}_{1:T}}$ where $p_\theta(\boldsymbol{x}_{0:T})$ is the joint distribution and $\boldsymbol{x}_1, ..., \boldsymbol{x}_T$ are latents. The data are perturbed by gradually introducing noise to $\boldsymbol{x}_0 \sim q(x_0)$, formalized by a Markov chain, known as forward process:

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \tag{2}$$

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{\alpha_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}) \tag{3}$$

where β_t is the noise schedule and $\alpha_t = 1 - \beta_t$. The diffusion model $p_\theta(\boldsymbol{x}_{t=1}|x_t)$ parameterized by θ is trained to approximate the reverse transition $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$, which is formulated as

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0), \tilde{\beta}_t \mathbf{I}) \tag{4}$$

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\boldsymbol{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t \tag{5}$$

where $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$ and $\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

The loss function of a noise prediction network is denoted as $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$:

$$\mathcal{L}_1(\theta) := \mathbb{E}_{t, \boldsymbol{x}_0, \boldsymbol{\epsilon}} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}, t)\|^2 \tag{6}$$

where t is uniform between 1 and T.

Then, denoising diffusion implicit model is learned to (DDIM) [12] generalize the framework of denoising diffusion probabilistic model (DDPM) [7] and propose a deterministic ODE process, achieving faster sampling speed.

At each diffusion step, given a noisy sample $\boldsymbol{x}_t$, a prediction of the noise-free sample $\hat{\boldsymbol{x}}_0$ along with a direction that points to $\boldsymbol{x}_t$ is computed. The final prediction of $\boldsymbol{x}_{t-1}$ is obtained by:

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\hat{\boldsymbol{x}}_0^t + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\epsilon_\theta(\boldsymbol{x}_t, t) + \sigma_t \epsilon_t \tag{7}$$

$$\hat{\boldsymbol{x}}_0^t = \frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{x}_t)}{\sqrt{\alpha_t}} \tag{8}$$

where α_t and σ_t are the parameters of the scheduler and ϵ_θ is the noise predicted by the decoder at the current step t.

Therefore, the denoised image frames $\boldsymbol{r}$ is denoted as $\hat{\boldsymbol{x}}_0^t$.

To get the corresponding background and foreground,

$$[\boldsymbol{m}_{fg}, \boldsymbol{m}_{bg}] = \mathcal{F}_1(\boldsymbol{r}) \tag{9}$$

where $\mathcal{F}_1$ is implemented by employ Segment Anything Model (SAM) [8] to get $\boldsymbol{m}_{fg}$ (foreground) and $\boldsymbol{m}_{bg}$ (background).

3.2. Average Fusion Module

The key objective of this module is to optimize the consistency of the image frames to make them have a smooth transition between frames and maintain the foreground area for future fine-tuning. Given the original image frames $\boldsymbol{r}$ and their masked areas $\boldsymbol{m}_{fg}$, $\boldsymbol{m}_{bg}$, we can obtain an optimal $\boldsymbol{r}'_k$ at certain step $t = k$ in the diffusion model:

$$\boldsymbol{r}'_k = \mathcal{F}_2(\boldsymbol{r}, \boldsymbol{m}_{fg}, \boldsymbol{m}_{bg}) \tag{10}$$

$\mathcal{F}_2$ is denoted by:

$$\boldsymbol{r}'_k = \frac{1}{n} \cdot w_1 \sum_{i=1}^{n} \boldsymbol{m}_{bg_{i,k}} + w_2 \cdot \boldsymbol{m}_{fg_k} \tag{11}$$

where $\boldsymbol{m}_{bg_{i,k}}$ is the ith masked background areas randomly sampled from $\boldsymbol{r}$ at step k; $\boldsymbol{m}_{fg_k}$ is the masked foreground area of $\boldsymbol{r}_k$ at step k, w_1, w_2 are the learnable weights for background and foreground. The flow diagram of the procedures is illustrated in figure 2.

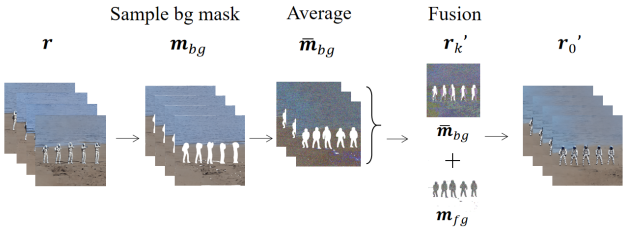

Figure 2. The flow diagram of the proposed average fusion method.

Then we conduct the remaining denoising procedures via equation (8) to get $\boldsymbol{r}'_0$ denoting as $\boldsymbol{r}'$:

$$\boldsymbol{r}' = \mathcal{F}_3(\boldsymbol{r}'_k, \boldsymbol{h}, \boldsymbol{c}) \tag{12}$$

where $\mathcal{F}_3$ is the function of the denoising model with modified $\boldsymbol{r}'_k$ at step k.

3.3. Combined Tuning Module

Afterward, a proposed fine-tune process to optimize the diffusion model parameters with $\boldsymbol{r}'$ and text embeddings $\boldsymbol{h}$ is conducted. This is achieved by the following equations:

$$\mathcal{D}' = \mathcal{F}_4(\mathcal{D}, \boldsymbol{r}', \boldsymbol{h}) \tag{13}$$

where $\mathcal{D}$ is the diffusion model with a denoising model $\boldsymbol{\epsilon_\theta}$ of parameter $\boldsymbol{\theta}$.

This procedure $\mathcal{F}_4$ can be formulated as an optimization problem:

$$\arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\|(\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,\boldsymbol{h}))\|^2 \quad (14)$$

To better utilize the properties of foreground and background, We apply separate weights to the foreground and background areas to finely tune the output video effects.

$$\mathcal{L}_{fg} = E_{\boldsymbol{x}_0,\boldsymbol{x}_t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\|M_{fg}\otimes(\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,\boldsymbol{h}))\|^2 \quad (15)$$

$$\mathcal{L}_{bg} = E_{\boldsymbol{x}_0,\boldsymbol{x}_t,\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\|M_{bg}\otimes(\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,t,\boldsymbol{h}))\|^2 \quad (16)$$

where $\otimes$ represent the Hadamard product, $\mathcal{L}_{fg}$ is the loss of foreground, $\mathcal{L}_{bg}$ is the loss of background, M_{fg} and M_{bg} are masks of foreground and background.

3.4. Inter-frame Consistency Module

Afterwards, the inter-frame consistency module [14] is proposed as follows:

$$\boldsymbol{v} = \mathcal{F}_5(\mathcal{D}', \boldsymbol{h}, \boldsymbol{c}) \quad (17)$$

$\mathcal{F}_5$ considers the denoising process of the entire video as multiple short videos with temporal overlapping undergoing parallel denoising in the temporal domain [14]. Suppose a model $\mathcal{D}^l_{\boldsymbol{\theta}}(\boldsymbol{v}_{t-1}|\boldsymbol{v}_t,\boldsymbol{h})$ in layer l with corresponding noise prediction network $\boldsymbol{\epsilon}^l_{\boldsymbol{\theta}}(\boldsymbol{v}_t,t,\boldsymbol{h})$ capable of denoising the given long video $\boldsymbol{v}_t$ such that $\boldsymbol{v}_{t-1}\sim\mathcal{D}^l_{\boldsymbol{\theta}}(\boldsymbol{v}_{t-1}|\boldsymbol{v}_t,\boldsymbol{h})$.

Define a mappings $\mathcal{P}_i$ which projects original videos $\boldsymbol{v}_t$ in the trajectory to short video segments $\boldsymbol{v}^i_t$:

$$\mathcal{P}_i(\boldsymbol{v}_t) = \boldsymbol{v}^i_t = \boldsymbol{v}_{t,S\times i:S\times i+K} \quad (18)$$

where $t = 1,2,...,T$, $i = 0,1,...,N-1$, $\boldsymbol{v}_{t,S\times i:S\times i+K}$ represents the collection of frames with sequence number from $S\times i$ to $S\times i+K$, S represents the stride among adjacent short video clips, K is the fixed length of short videos, N is the total number of clips. Each short video $\boldsymbol{v}^i_t$ is guided with an independent text condition $\boldsymbol{h}^i$.

For simplicity, the diffusion models for all video clips are set to one single model $\mathcal{D}'_{\boldsymbol{\theta}}(\boldsymbol{v}^i_{t-1}|\boldsymbol{v}^i_t,\boldsymbol{h}^i)$.

The optimal $\boldsymbol{v}_{t-1}$ can finally be obtained by solving the following optimization problem:

$$\boldsymbol{v}_{t-1} = \arg\min_{\boldsymbol{v}} \sum_{i=0}^{N-1}\|W_i\otimes(\mathcal{P}_i(\boldsymbol{v})-\boldsymbol{v}^i_{t-1})\|^2 \quad (19)$$

where W_i is the pixel-wise weight for the video clip $\boldsymbol{v}^i_t$, $\otimes$ is the Hadamard product. For an arbitrary frame j in the video $\boldsymbol{v}_{t-1}$, it is equal to the weighted sum of all the corresponding frames in short videos that contain the j frame [14].

3.5. Training objective

Therefore, the total loss $\mathcal{L}_{total}$ for the optimization problem of generating the videos is defined as follows:

$$\mathcal{L}_{total} = \lambda_1\cdot\mathcal{L}_1 + \lambda_2\cdot\mathcal{L}_{fg} + \lambda_3\cdot\mathcal{L}_{bg} \quad (20)$$

where $\mathcal{L}_1, \mathcal{L}_{fg}, \mathcal{L}_{bg}$ are loss functions derived from equation (6),(15),(16) and $\lambda_1,\lambda_2,\lambda_3$ are their corresponding weights.

4. Experiments

Our process involves editing of videos generated based on text. The implementation details of the proposed framework for video generation is shown as follows:

- Input the text prompt and conditional pose frames to generate the corresponding embedding, then load the stable diffusion model v1-4 [10] and T2I-Adapter [9] model for image generation. The β in the DDPM noise scheduler starts at 8.5e-4 and ends at 1.2e-2, with the scaled linear method and total diffusion steps of 1000.
- Randomly sample 5 results at 5th step from $\boldsymbol{r}_{995}$ at the back process of DDPM and get $\boldsymbol{r}_0$, apply SAM to $\boldsymbol{r}_0$ to obtain the foreground $\boldsymbol{m}_{fg}$ and background $\boldsymbol{m}_{bg}$, create $\boldsymbol{r}'_{995}$ by equation (11) and estimate $\boldsymbol{r}'_0$ by $\boldsymbol{r}'_{995}$ via back process.
- Fine tune the diffusion model with learning rate of 2e-6 for 250 steps and batch size 1 with a modified loss function by equation (20) and output $\mathcal{D}'$ in equation (13).
- Process the inter-frame consistency module with the input of a modified diffusion model $\mathcal{D}'$, text embeddings $\boldsymbol{h}$, condition $\boldsymbol{c}$ and output the final long video $\boldsymbol{v}$ (see equation (17)).

5. Result

Currently, we apply the first and second module (separate tuning module, average fusion module) and produce the initial result. We provide a visual presentation of our results in Figure 3, in comparison with the state-of-the-art [14] and our proposed framework. It can be seen our method generates videos with a good consistency. We are going to apply the left modules and produce the final results.

6. Conclusion

We propose a video generation framework with four modules to generate a long video with good consistency. We complete our first experiment applying the separate tuning module and average fusion module, showing the experiment in comparison with the state-of-the-art. Then, we are going to apply the remaining modules and complete our left experiments.

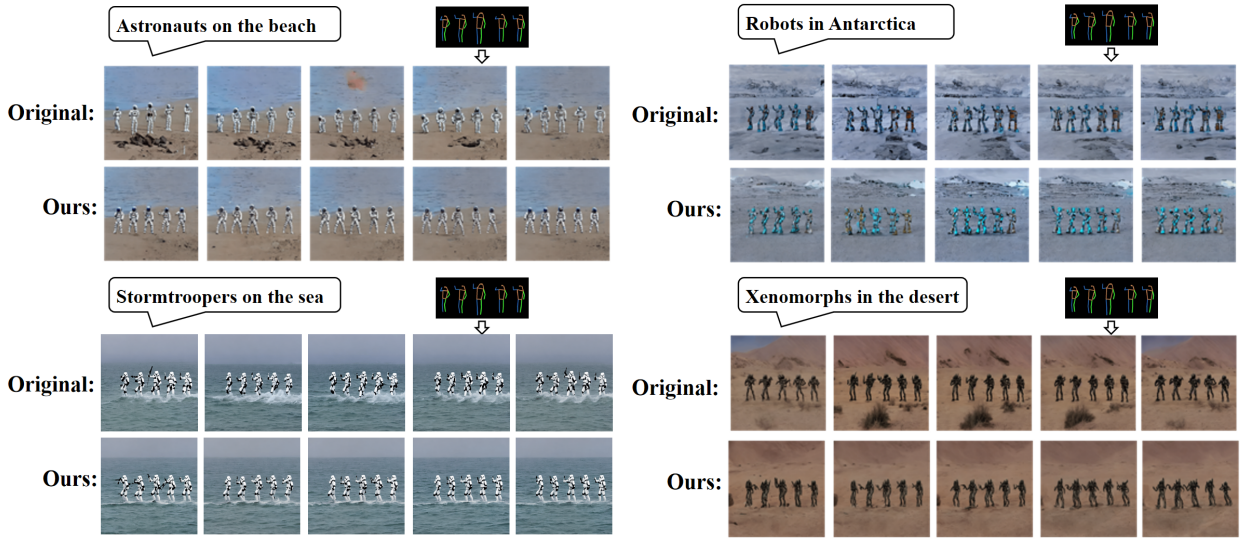

Figure 3. Comparison of frames between the state-of-the-art and our proposed framework.

References

[1] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. *arXiv preprint arXiv:2303.12688*, 2023.

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[3] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022.

[4] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022.

[5] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022.

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

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

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[9] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

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

[11] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[13] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 35:23371–23385, 2022.

[14] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023.

[15] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*, 2022.

[16] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18456–18466, 2023.