

000
001
002
003
004
005
006
007
008
009
010
011

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

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

Anonymous ICCV submission

Paper ID 12

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.

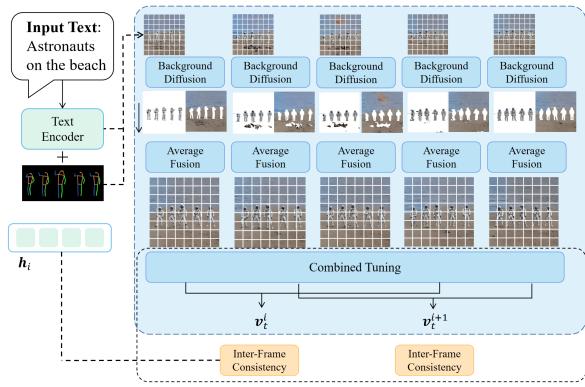


Figure 1. A high level overview of the proposed video generation approach.

3.1. Separate Tuning Module

Given the input prompts of the hidden embeddings h , and the corresponding condition c , image frames r are cre-

108 ated by

$$\mathbf{r} = \mathcal{F}_0(\mathbf{h}, \mathbf{c}) \quad (1)$$

112 where \mathcal{F}_0 denotes the function of the diffusion and denoising
113 process.

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

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

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

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

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

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

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

136 The loss function of a noise prediction network is de-
137 noted as $\epsilon_\theta(\mathbf{x}_t, t)$:

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

140 where t is uniform between 1 and T .

141 Then, denoising diffusion implicit model is learned to
142 generalize the framework of denoising diffusion
143 probabilistic model (DDPM) [7] and propose a deter-
144 ministic ODE process, achieving faster sampling speed.

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

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

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

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

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

159 To get the corresponding background and foreground,

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

162 where \mathcal{F}_1 is implemented by employ Segment Anything
163 Model (SAM) [8] to get \mathbf{m}_{fg} (foreground) and \mathbf{m}_{bg} (back-
164 ground).

3.2. Average Fusion Module

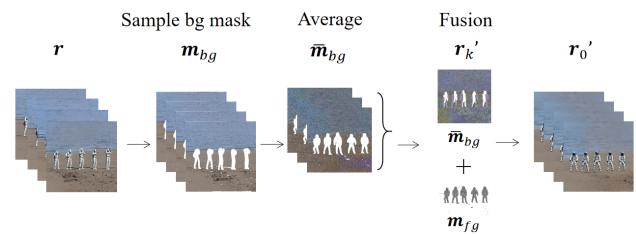
166 The key objective of this module is to optimize the con-
167 sistency of the image frames to make them have a smooth
168 transition between frames and maintain the foreground area
169 for future fine-tuning. Given the original image frames \mathbf{r}
170 and their masked areas \mathbf{m}_{fg} , \mathbf{m}_{bg}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

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

174 \mathcal{F}_2 is denoted by:

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

178 where $\mathbf{m}_{bg_{i,k}}$ is the i th masked background areas randomly
179 sampled from \mathbf{r} at step k ; \mathbf{m}_{fg_k} is the masked foreground
180 area of \mathbf{r}_k at step k , w_1, w_2 are the learnable weights for
181 background and foreground. The flow diagram of the pro-
182 cedures is illustrated in figure 2.



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

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

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

203 where \mathcal{F}_3 is the function of the denoising model with mod-
204 ified \mathbf{r}'_k at step k .

3.3. Combined Tuning Module

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

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

212 where \mathcal{D} is the diffusion model with a denoising model ϵ_θ
213 of parameter θ .

216 This procedure \mathcal{F}_4 can be formulated as an optimization
 217 problem:
 218

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

221 To better utilize the properties of foreground and background,
 222 We apply separate weights to the foreground and background areas to finely tune the output video effects.
 223

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

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

229 where \otimes represent the Hadamard product, \mathcal{L}_{fg} is the loss
 230 of foreground, \mathcal{L}_{bg} is the loss of background, M_{fg} and M_{bg}
 231 are masks of foreground and background.
 232

233 3.4. Inter-frame Consistency Module

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

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

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

246 Define a mappings \mathcal{P}_i which projects original videos \mathbf{v}_t
 247 in the trajectory to short video segments \mathbf{v}_t^i :

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

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

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

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

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

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

270 3.5. Training objective

271 Therefore, the total loss \mathcal{L}_{total} for the optimization prob-
 272 lem of generating the videos is defined as follows:

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

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

276 4. Experiments

277 Our process involves editing of videos generated based
 278 on text. The implementation details of the proposed frame-
 279 work for video generation is shown as follows:

- 280 • Input the text prompt and conditional pose frames to gen-
 281 erate the corresponding embedding, then load the stable
 282 diffusion model v1-4 [10] and T2I-Adapter [9] model for
 283 image generation. The β in the DDPM noise scheduler
 284 starts at 8.5e-4 and ends at 1.2e-2, with the scaled linear
 285 method and total diffusion steps of 1000.
- 286 • Randomly sample 5 results at 5th step from r_{995} at the
 287 back process of DDPM and get r_0 , apply SAM to r_0 to
 288 obtain the foreground \mathbf{m}_{fg} and background \mathbf{m}_{bg} , create
 289 r'_{995} by equation (11) and estimate r'_0 by r'_{995} via back
 290 process.
- 291 • Train the proposed model with learning rate of 2e-6 for
 292 250 steps and batch size 1 and tune the weights of sepa-
 293 rate tuning module, average fusion module and combined
 294 tuning module and output \mathcal{D}' (see equation (17)).
- 295 • Process the inter-frame consistency module with the in-
 296 put of a modified diffusion model \mathcal{D}' , text embeddings \mathbf{h} ,
 297 condition \mathbf{c} and output the final long video \mathbf{v} .

298 5. Result

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

311 6. Conclusion

312 We propose a video generation framework with four
 313 modules to generate a long video with good consistency.
 314 We complete our first experiment applying the separate tun-
 315 ing module and average fusion module, showing the exper-
 316 iment in comparison with the state-of-the-art. Then, we are
 317 going to apply the next three modules and complete our left
 318 experiments.

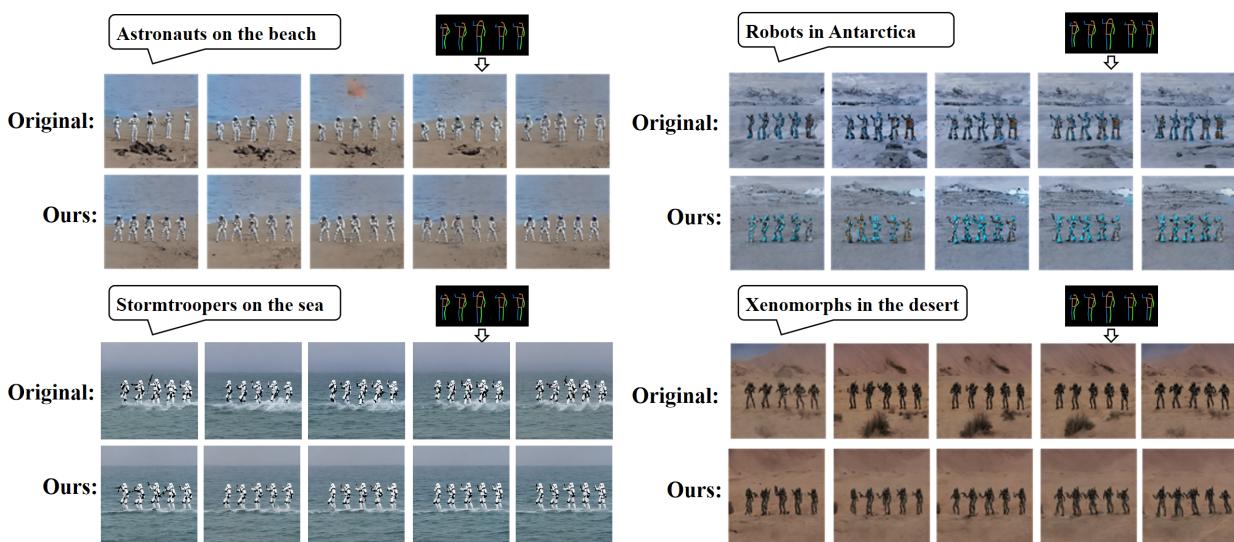


Figure 3. Comparison of frames between the state-of-the-art and our proposed framework.

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, Zhonggang 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.