FasterVD: On Acceleration of Video Diffusion Models

Pinrui Yu[1] , **Dan Luo**[1] , **Timothy Rupprecht**[1] , **Lei Lu**[1] , **Zhenglun Kong**[1] , **Pu Zhao**[1] ,
Yanyu Li[1] , **Octavia Camps**[1] , **Xue Lin**[1] and **Yanzhi Wang**[1]
[1]Northeastern University
{yu.pin, luo.dan1, rupprecht.t, lu.lei1, kong.zhe, p.zhao, li.yanyu, o.camps, xue.lin,
yanz.wang}@northeastern.edu

Abstract

Equipped with Denoising Diffusion Probabilistic Models, video content generation has gained significant research interest recently. However, diffusion pipelines call for intensive computation and model storage, which poses challenges for their wide and efficient deployment. In this work, we address this issue by integrating LCM-LoRA to reduce the denoising steps and escalating the video generation process by frame skipping and interpolation. Our framework achieves an approximately 10× inference acceleration for high-quality realistic video generation on commonly available GPUs.

1 Introduction

The domain of image & video generation has made remarkable advancements in generative capability and quality with the advent of diffusion models [Song *et al.*, 2020b; Guo *et al.*, 2023; Hu *et al.*, 2023; Blattmann *et al.*, 2023a; Girdhar *et al.*, 2023]. In this work, we investigate the challenging video generation task, where Magic-Animate [Xu *et al.*, 2023a] stands out with its ability to animate static human images to high-quality videos by applying specific motion sequences, serving as a strong baseline paradigm.

Despite the promising performance, it is challenging to enable an efficient video generation pipeline for wide deployments due to the following: (i) High cost for multiple steps and frames. Diffusion models require tens of denoising steps, with 10^2 to 10^3 GMACs of computation for each step. Furthermore, the generated video may consist of more than hundreds of frames to ensure smooth and informative animations, incurring substantial computational and memory costs. (ii) Complicated temporal dynamics. To model the diverse visual contents across frames with complicated temporal dynamics, video diffusion models typically employ extra temporal convolution and attention layers [Luo *et al.*, 2023c; He *et al.*, 2022], which further amplify the computation costs. (iii) Quality and efficiency trade-off. To alleviate these burdens, recent works propose to reduce the denoising steps [Salimans and Ho, 2022; Li *et al.*, 2024; Luo *et al.*, 2023a] or compress the network [Li *et al.*, 2024; Kim *et al.*, 2023; Zhao *et al.*, 2023]. However, these methods may suffer from substantial performance degradation in video generation.

The ubiquitous mobile devices are ideal for interactive applications such as personalized digital emoji generation [Bai *et al.*, 2019] and human image animation [Xu *et al.*, 2023b]. However, due to the tremendous parameters and computations introduced by diffusion-based generative models, the inference latency for generating a several-second video, even with powerful GPUs, would be measured in minutes. Consequently, such dilemma becomes more formidable when confronted with computational resource constraints on mobile devices. Thus, it is of great necessity to build efficient video generation pipelines capable of real-time deployment across various platforms, particularly on mobile devices.

To address these challenges, our work strategically incorporates a pre-trained LCM-LoRA [Luo *et al.*, 2023b] module within Magic-Animate, augmented by a relatively computational-cheap yet effective frame insertion technique. The incorporation with LCM-LoRA not only remarkably reduces the computational demands with fewer denoising steps (from 25 to 4 with 6× speedup), but also meticulously preserves the high-quality outputs. Furthermore, we propose a frame insertion method capable of generating key frames exclusively through the diffusion model. We then bridge these key frames with a lightweight frame interpolation module, thus reducing the costs of frame generation. Our superior performance in terms of inference efficiency and generation quality paves the way for advanced real-time animation deployment on mobile platforms. Our contributions can be concluded as:

- We offer a novel framework by incorporating the LCM-LoRA and frame insertion with Magic-Animate to significantly accelerate the inference speed with about 10× speedup for high-quality video generation.
- To the best of our knowledge, we are the first to deploy the image-to-video diffusion-based animation models on mobile devices, demonstrating the potential for generating high-quality, real-time animated video on the edge.

2 Diffusion-Based Video Generation

Denoising Diffusion Probabilistic Model learns a reverse function to predict real data distribution $\mathbf{x} \sim p_{\text{data}}$ from its noisy version $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$, where α_t and σ_t are determined by the selected noise scheduler. Specifically, given a denoising U-Net [Ronneberger *et al.*, 2015; Ho *et al.*, 2020;

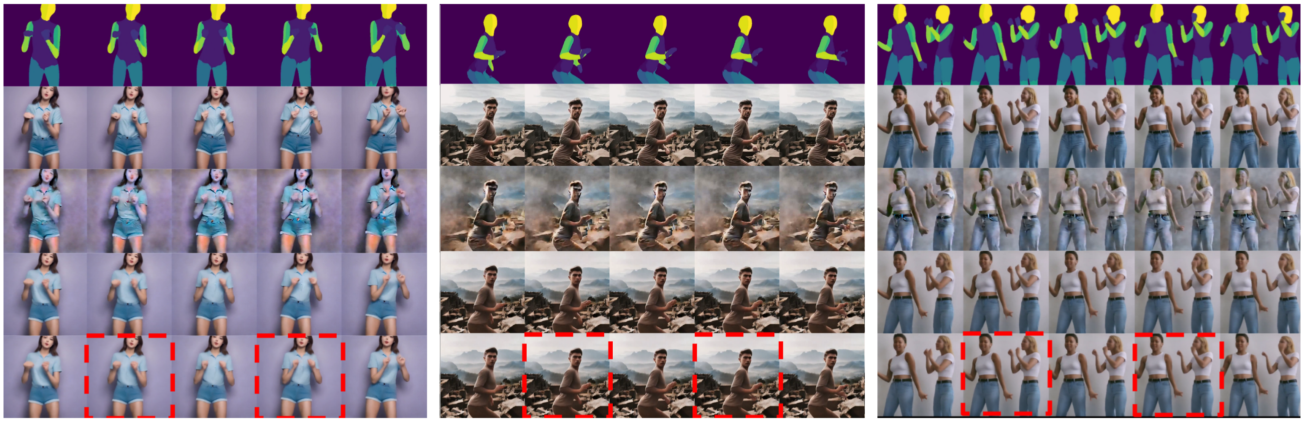

Figure 1: Visualization comparison between our proposed LCM-LoRA Magic-Animate with frame insertion and the original Magic-Animate model. The rows from top to bottom are the dense pose, 25 step Magic Animate, 4 step Magic Animate, 4 step Magic Animate with LCM-LoRA, and then finally our 4 step Magic Animate with LCM-LoRA and interpolation by key frame. Frames in red box are generated by key frames.

Karras *et al.*, 2022; Rombach *et al.*, 2022] model $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\cdot)$ parameterized by $\boldsymbol{\theta}$, the training objective can be formulated as follows [Sohl-Dickstein *et al.*, 2015; Song *et al.*, 2020b]:

$$\min_{\boldsymbol{\theta}} \mathbf{E}_{t\sim p_t, \mathbf{x}\sim p_{\text{data}}, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \, ||\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(t, \mathbf{z}_t) - \epsilon||_2^2, \qquad (1)$$

where t is the time step and ϵ is the ground-truth noise. A converged model $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ can then iteratively revert the diffusion process and generate $\mathbf{z}_0 = \mathbf{x}$ from noisy $\mathbf{z}_T$ with various schedulers. Taking the first-order DDIM [Song *et al.*, 2020a] as an example,

$$\mathbf{z}_{t-1} = \alpha_{t-1}\frac{\mathbf{z}_t - \sigma_t \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(t, \mathbf{z}_t)}{\alpha_t} + \sigma_{t-1}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(t, \mathbf{z}_t), \qquad (2)$$

where the process is repeated until we obtain $\mathbf{z}_0$.

From Image to Video Generation. In this work, we investigate video generation task, extending image diffusion models ($\mathbf{x} \in \mathbb{R}^{\mathtt{b,c,h,w}}$) to generate a sequence of frames ($\mathbf{x} \in \mathbb{R}^{\mathtt{b,f,c,h,w}}$). Spatial modeling is formulated identically to image generation by parallelizing the batch and frame dimension (*i.e.,* $\mathtt{b}\cdot\mathtt{f},\mathtt{c},\mathtt{h},\mathtt{w}$). Additionally, 3D convolutions ($\mathtt{b},\mathtt{c},\mathtt{f},\mathtt{h},\mathtt{w}$) and temporal attentions ($\mathtt{b},\mathtt{h}\cdot\mathtt{w},\mathtt{f},\mathtt{c}$) are introduced to model temporal dependencies. By feeding different input conditions to the U-Net model $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\cdot)$, one can formulate different types of video generation, such as text-to-video [Guo *et al.*, 2023; Blattmann *et al.*, 2023b], image-to-video [Blattmann *et al.*, 2023a], motion-to-video [Lin *et al.*, 2022; Hu *et al.*, 2023], etc.

3 Method

Diffusion Model Acceleration. To mitigate the computation and storage demands of Diffusion Models, various strategies have been proposed. Among them, LCM-LoRA [Luo *et al.*, 2023b] has gained popularity for its ability to expedite the model inference without the need for model-specific training. This method involves integrating LoRA distillation into pre-trained Stable-Diffusion models, coupled with identifying LoRA parameters through a process known as LCM distillation. However, it is only used in image generation, and cannot be directly applied to video.

3.1 LCM-LoRA Step Reduction

When directly plugging the pretrained LCM-LoRA into the corresponding stable diffusion models to reduce the denoising steps, we need to pay special attention to the classifier-free guidance scale for video generation. Specifically, the guidance scale controls the influence of the input prompt on the generated image. A higher value of the guidance scale increases the model's adherence to the input prompt, resulting in images more closely matching the specified instructions with less freedom to generate noises in each step. In normal text2image and image2image Stable-Diffusion scenarios, there is no need for specific temporal guidance when generating one single image. However, in human image animation with video outputs instead of a single static image,the generation of multiple frames requires temporal guidance from the previous frame, which enforces the tuning of the guidance scale with the LCM-LoRA setting. We need to decrease the guidance scale to give the model more creative freedom with far fewer generation steps.

Framework Compatibility. LCM-LoRA, when applied to Stable Diffusion models, specifically targets the U-Net architecture, which is a core component of the model responsible for the single image generation process and also particularly suited for further optimizations. In the Magic-Animate, the authors build a new temporal 3D U-Net to not only generate high-quality multi-frame images but also keep temporal coherence across frames, ensuring that the generated animations are smooth and consistent over time, without abrupt changes that would break the illusion of motion. Compared to the original U-Net of stable diffusion models, the Magic-Animate's 3D temporal U-Net involves 3D temporal consistency layers and dynamic attention blocks. Thus, we selectively integrate the LCM-LoRA parameters to the corresponding part of the new 3D U-Net while keeping the remaining parameters from Magic-Animate unchanged, and re-tune the original training parameters to make sure a high generative quality of human animation video.

Magic-Animate	Params (M)	GFLOPs	# Forward	iPhone Latency (ms)	GPU Latency (ms)	Total GPU Latency (s)	All GPU Latency (s)
Text Encoder	123	9	1	4	2	0.002	32.5
Appearance Encoder	857	65	25	206	32	0.8	
ControlNet	857	41	25	192	17	0.4	
U-Net	1368	590	25x2	4146	603	30.2	
VAE Decoder	50	1240	16	369	67	1.1	

Our Model	Params (M)	GFLOPs	# Forward	iPhone Latency (ms)	GPU Latency (ms)	Total GPU Latency (s)	All GPU Latency (s)
Text Encoder	123	9	1	4	2	0.002	**3.5**
Appearance Encoder	857	65	**4**	206	32	**0.13**	
ControlNet	857	41	**4**	192	17	**0.07**	
U-Net	1368	590	**4x2**	**2073**	**302**	**2.4**	
VAE Decoder	50	1240	**16/2**	369	67	**0.54**	
Frame Interpolation Module	34	558	**16/2**	246	42	0.34	

Table 1: Latency comparison between original Magic-Animate and our proposed LCM-LoRA Magic-Animate with frame insertion model on NVIDIA Titan RTX 24G GPU and iPhone 14 Pro Max. Here all results are based on generating a 1-second video with 16fps. We use half frame interpolation (8 frames) as an example here. In # Forward, the U-Net "4x2" indicates 4 stable diffusion steps and batch size of 2 in a video sample. Our model achieves half the latency of the U-Net because it only needs to generate only key frames instead of a full video.

3.2 Frame Insertion

Human animation video generation methods [Guo *et al.*, 2023; Xu *et al.*, 2023a] generate a series of continuous frames via a unified denoising diffusion process. The iterative denoising together with frame-based temporal attention blocks to ensure inter-frame consistency have introduced high computational workload, making it infeasible for edge implementation. Such burden motivates us towards generating only intermittent frames from the expected output, and bridging the inter-frame gaps through a computationally cost-effective frame interpolation algorithm. We propose a key-frame based generation algorithm, proportionally reducing the generation latency for the frame sequence. Moreover, a SOTA frame interpolation module [Reda *et al.*, 2022] is integrated for intermediate-frame insertion between generated key frames, ensuring the perceived continuity of the motion sequence.

Key Pose Selection. In Magic-Animate, due to the strong one-to-one frame correspondence between the output video and the DensePose input motion sequence (i.e., original generated frames $I_{1:\mathrm{T}}$ align with original full Densepose sequence $P_{1:\mathrm{T}}$), the Key Pose Selection $F(P_{1:\mathrm{T}}) = \{P_1,P_a,P_b,...,P_T\}$ is vital for correlated key-frame generation. Thus, rather than prefixing a selection stride, we develop a simple yet effective frame-likelihood based key-pose selection algorithm as the unified approach. Specifically, following a selected key-pose frame P_a, we determine its sequential frame P_{a+i} (i starts from 1) as the next key frame if $\Delta_{a,a+i} \geq \lambda_{\mathrm{MSE}}$, where λ_{MSE} is a hyperparameter as the frame-likelihood threshold and $\Delta_{a,a+i}$ denotes the MSE between P_a and P_{a+i}. If $\Delta_{a,a+i} < \lambda_{\mathrm{MSE}}$, we continue the search by $i = i + 1$ until the next key frame is found. Then we start another round of search by setting $a = a + i$. The first and last pose frames P_1 and P_T are pre-selected for every key frame inputs, ensuring the general alignment of the selected motion guidance after dropping the original counterpart. With such selection schedule, highly close or similar motion frames are removed, while the transition points within the motion sequence are retained, thereby extracting the key information of the entire sequence. Moreover, the adjustable λ_{MSE} achieves a straight control towards the trade-off between generation quality and latency.

Key Frame Generation. In consistency with the selected key-pose frames $P_{\mathrm{select}} = \{P_1,P_a,P_b,...,P_T\}$ serving as the guidance input for Magic-Animate, the temporal embeddings, which depict the chronological order for the noisy frame inputs in the temporal domain, are also adjusted correspondingly. The selected key-pose sequence P_{select} is directly fed into Magic-Animate as the control signal. Thus, the computation of the iterative-denoising diffusion model would be greatly shrunk with the proper setting of the motion-dropping threshold λ_{MSE}, while the generation quality would be well maintained without structural modifications on the temporal-related blocks within the diffusion model.

4 Experiments and Visualization

Experiments. In Table 1, we demonstrate the computation cost (GFLOPs) and latency of our LCM-LoRA Magic-Animate with frame insertion on both the NVIDIA Titan RTX GPU and the iPhone 14 Pro Max, with half-precision (FP16). To perform a comprehensive study, we compare with each part of the original Magic-Animate model. As observed, by leveraging LCM-LoRA to reduce the original 25 denoising steps to just 4 steps, along with frame insertion to halve the number of generated frames from the computationally expensive U-Net, our model achieves approximately 10× inference speedup and preserves generative quality. With our efficient design, the implementation of human animation video generation on mobile devices can be realized within a reasonable timeframe.

Visualization. In Figure 1, we compare the visualization results between video frames generated by the original Magic-Animate and our LCM-LoRA Magic-Animate with frame insertion. The introduction of LCM-LoRA into the Magic-Animate reduces the required denoising steps from the original 25 to an impressive 4, while frame insertion further halves the frame number from the U-Net diffusion process, without compromising the quality of the output. To make a fair comparison, we show the 4-step results of the original Magic-Animate. We can clearly observe that the characters in these videos begin to distort, with the surrounding background showing a lot of noise and chaotic blur.

Acknowledgments

This research was partially supported by National Science Foundation IIS-2310254, NSF grant 2038493, ONR grant N00014-21-1-2431, and U.S. Department of Homeland Security grant 22STESE00001-03-02.

Contribution Statement

Pinrui Yu, Dan Luo, and Timothy Rupprecht contributed equally to this work as first authors.

References

[Bai *et al.*, 2019] Qiyu Bai, Qi Dan, Zhe Mu, and Maokun Yang. A systematic review of emoji: Current research and future perspectives. *Frontiers in psychology*, 10:2221, 2019.

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

[Blattmann *et al.*, 2023b] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[Girdhar *et al.*, 2023] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

[Guo *et al.*, 2023] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[He *et al.*, 2022] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[Hu *et al.*, 2023] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.

[Karras *et al.*, 2022] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

[Kim *et al.*, 2023] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. On architectural compression of text-to-image diffusion models. *arXiv preprint arXiv:2305.15798*, 2023.

[Li *et al.*, 2024] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *Advances in Neural Information Processing Systems*, 36, 2024.

[Lin *et al.*, 2022] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022.

[Luo *et al.*, 2023a] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[Luo *et al.*, 2023b] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023.

[Luo *et al.*, 2023c] Z. Luo, D. Chen, Y. Zhang, Y. Huang, L. Wang, Y. Shen, D. Zhao, J. Zhou, and T. Tan. Notice of removal: Videofusion: Decomposed diffusion models for high-quality video generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10209–10218, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society.

[Reda *et al.*, 2022] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision*, pages 250–266. Springer, 2022.

[Rombach *et al.*, 2022] 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.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICAI*, 2015.

[Salimans and Ho, 2022] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

[Sohl-Dickstein *et al.*, 2015] 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.

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

[Song *et al.*, 2020b] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[Xu *et al.*, 2023a] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023.

[Xu *et al.*, 2023b] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. 2023.

[Zhao *et al.*, 2023] Yang Zhao, Yanwu Xu, Zhisheng Xiao, and Tingbo Hou. Mobilediffusion: Subsecond text-to-image generation on mobile devices. *arXiv preprint arXiv:2311.16567*, 2023.