FasterVD: On Acceleration of Video Diffusion Models

Pinrui Yu\textsuperscript{1}, Dan Luo\textsuperscript{1}, Timothy Rupprecht\textsuperscript{1}, Lei Lu\textsuperscript{1}, Zhenglun Kong\textsuperscript{1}, Pu Zhao\textsuperscript{1}, Yanyu Li\textsuperscript{1}, Octavia Camps\textsuperscript{1}, Xue Lin\textsuperscript{1} and Yanzhi Wang\textsuperscript{2}

\textsuperscript{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 $10\times$ 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 $6\times$ 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 $10\times$ 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 $x \sim p_{\text{data}}$ from its noisy version $z_t = \alpha_t x + \sigma_t \epsilon$, where $\alpha_t$ and $\sigma_t$ are determined by the selected noise scheduler. Specifically, given a denoising U-Net [Ronneberger et al., 2015; Ho et al., 2020;
where the process is repeated until we obtain the converged model $\epsilon^*$. The training objective can be formulated as follows [Sohl-Dickstein et al., 2015; Song et al., 2020b]:

$$\min_\theta \mathbb{E}_{t \sim p_t, x \sim p_{data}, \epsilon \sim N(0,1)} \|\epsilon \theta(t, z_t) - \epsilon\|^2_2,$$  \hspace{1cm} (1)

where $t$ is the time step and $\epsilon$ is the ground-truth noise. A converged model $\epsilon \theta$ can then iteratively revert the diffusion process and generate $z_0 = x$ from noisy $z_T$ with various schedulers. Taking the first-order DDIM [Song et al., 2020a] as an example,

$$z_{t-1} = \alpha_{t-1} \frac{z_t - \sigma_t \epsilon \theta(t, z_t)}{\alpha_t} + \sigma_{t-1} \epsilon \theta(t, z_t),$$  \hspace{1cm} (2)

where the process is repeated until we obtain $z_0$.

### From Image to Video Generation

The authors develop the 4-step Magic-Animate with LCM-LoRA and interpolation by key frame.

#### 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.
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 $I_{t:T}$ align with original full DensePose sequence $P_{t:T}$), the Key Pose Selection $F(P_{t:T}) = \{P_i,P_{i+1},P_{i+2},...,P_f\}$ 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 $P_{a+1}$ (i starts from 1) as the next key frame if $\Delta_{a,a+1} \geq \lambda_{\text{MSE}}$, where $\lambda_{\text{MSE}}$ is a hyperparameter as the frame-likelihood threshold and $\Delta_{a,a+1}$ denotes the MSE between $P_a$ and $P_{a+1}$. If $\Delta_{a,a+1} < \lambda_{\text{MSE}}$, 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 $P_f$ 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 $\lambda_{\text{MSE}}$ achieves a straight control towards the trade-off between generation quality and latency.

Key Frame Generation. In consistency with the selected key-pose frames $P_{\text{select}} = \{P_i,P_{i+1},P_{i+2},...,P_f\}$ 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 $P_{\text{select}}$ 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 $\lambda_{\text{MSE}}$, 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.

### 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.

<table>
<thead>
<tr>
<th></th>
<th>Params (M)</th>
<th>GFLOPs (M)</th>
<th># Forward</th>
<th>iPhone Latency (ms)</th>
<th>GPU Latency (ms)</th>
<th>Total GPU Latency (s)</th>
<th>All GPU Latency (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text Encoder</strong></td>
<td>123 9</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>0.002</td>
<td>32.5</td>
</tr>
<tr>
<td><strong>Appearance Encoder</strong></td>
<td>857 65</td>
<td>25</td>
<td>25</td>
<td>206</td>
<td>32</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td><strong>ControlNet</strong></td>
<td>857 41</td>
<td>25</td>
<td>25</td>
<td>192</td>
<td>17</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td><strong>U-Net</strong></td>
<td>1368 590</td>
<td>25x2</td>
<td></td>
<td>4146</td>
<td>603</td>
<td>30.2</td>
<td></td>
</tr>
<tr>
<td><strong>VAE Decoder</strong></td>
<td>50 1240</td>
<td>16</td>
<td></td>
<td>369</td>
<td>67</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td><strong>Frame Interpolation Module</strong></td>
<td>34 558</td>
<td>16/2</td>
<td></td>
<td>246</td>
<td>42</td>
<td>0.34</td>
<td>3.5</td>
</tr>
</tbody>
</table>

...
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


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


