A study of video frame interpolation based on pre-trained video diffusion models using parameter-efficient fine-tuning

Luoxu Jin† Hiroshi Watanabe† Takeshi Chujoh‡ Zheming Fan‡
Sujun Hong‡ Ryo Ishimoto‡ Kakeru Sugimoto‡ Tomohiro Ikai‡
†Waseda University ‡ Sharp Corporation

Abstract: In this study, we investigate the use of state-of-the-art video generation models applied to the task of video frame interpolation. We propose to train the Parameter-Efficient Fine-Tuning (PEFT) module to learn video frame interpolation, and benefiting from the PEFT module fine tuning method, our approach avoids destroy the video generation capability of the base model and converges quickly after a few learning steps.

1 Introduction

Video frame interpolation is the process of generating intermediate frames between existing frames to increase the frame rate and smooth the motion in a video. Traditional interpolation methods, such as linear interpolation or optical flow-based techniques, often struggle with handling complex motion or producing realistic transitions.

With the introduction of generative models, video frame interpolation has been significantly improved. In this study, we investigate the feasibility of applying video generation models trained on large-scale video datasets to the task of video frame interpolation. We use Stable Video Diffusion[1] (SVD) as the base model and train PEFT modules specifically to learn frame interpolation.

2 Related Work

2.1 Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) is a set of techniques designed to adapt pre-trained models to specific tasks while minimizing the number of parameters that need to be updated. Instead of fine-tuning all model parameters, PEFT methods typically involve freezing most of the model and adjusting only a small subset of parameters, such as adapter layers or specific weights. This approach reduces computational costs and training time, making it feasible to use large models in resource-constrained environments.

3 Method

3.1 Condition Encoder

In the text-to-image domain, ControlNet[2] demonstrates that it is possible to control the generation of images through a variety of different control signals, therefore we refer to ControlNet structure. The trainable PEFT module copies the Encoder part of UNet and inherits the pre-trained weights. In practice, for N frames, we preserve only the first and the last frame, and replace the intermediate frames as zero images to obtain c. Also we use mask $m \in \{0,1\}^{h*w}$ with the same shape to concatenate to the conditional c in the channel dimension, where $m = 1$ indicating a conditional frame, and $m = 0$ indicating an unconditional frame. Because of the PEFT module copies the Encoder of the base model, the final output of the PEFT module has the same dimension as the Encoder output and can be directly summed into the Decoder shown in Figure 1. For training, we use commonly diffusion model noise prediction loss in Equation 1.

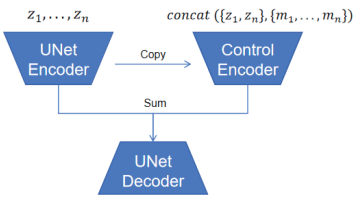

Figure 1: Condition Encoder Structure

$$L = E_{\epsilon\sim N(0,I),t\sim Uniform(1,T)}\,||\epsilon - \epsilon_\theta(z_t,t,C)|| \quad (1)$$

Here z_t represents a series of video frames after adding noise, t represents the current timestamp of the diffusion process, and C represents the output of the condition encoder.

3.2 Keyframe Attention

In order for the model to better learn the spatial and temporal correlation of conditional frames, we propose a simple attention mechanism. In the Stable Video Diffusion model, there are both spatial transformer blocks and temporal transformer blocks. We introduce a keyframe attention block within each of these blocks. Given the latent feature z, for N frames we have $\{z_1,...,z_n\}$, we derive query feature from frame z_i , key and value features from the first frame z_1 and the last frame z_n , and implement $Attention(Q,K,V)$ with follow equation:

$$Q = W^Q z_i, K = W^K[z_1,z_n], V = W^V[z_1,z_n], \quad (2)$$

Here we concatenate $[z_1,z_n]$ together in channel dimension for computation. Notice that our keyframe attention mechanism is at the beginning of each transformer block.

4 Experiment

4.1 qualitative results

In this section, we report the experimental results of video frame interpolation. During training, we fine-tuned the model using the Vimeo90K[3] dataset. Since the

Figure 2: Interpolation example in key frames with small motion

Figure 3: Interpolation example in key frames with large motion

dataset contains only 7-frame sequences, we provide the first and last frames during inference and predict the 5 intermediate frames. From the qualitative results below, we observe that given two conditional frames, the model successfully infers the movement in the middle frame. Moreover, our base model benefits from the learned video generation priors, allowing it to reasonably predict intermediate frames even when there are nonlinear changes and large motions between two frames shown in Figure 2 and Figure 3.

4.2 quantitative results

In this section, we report quantitative results and the experiments will evaluate the model video frame interpolation performance in terms of SSIM,PSNR and LPIPS metrics on DAVIS-7[4] shown in Table 1. Notice that the model gives good results even when predicting the middle 5 frames during training and 7 frames during inference.

	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	19.17	0.4603	0.3258
RIFE[5]	19.54	0.4546	0.2954
FILM[6]	19.75	0.4718	0.3048
LDMVFI[7]	19.07	0.4175	0.2765
VIDIM[8]	18.73	0.4221	0.2986

Table 1: Comparison of different video interpolation baselines on reconstruction and generation metrics, evaluating intermediate frames out of all 7 generated frames.

Our video interpolation model performs similarly to recent video interpolation models. On the DAVIS dataset, which contains large-motion sequences, our model is able to predict the intermediate motion frames. Compared to models that are limited to linear interpolation, our model can generate more plausible intermediate frames.

5 Conclusion

We propose adopting the PEFT method to fine-tune the Stable Video Diffusion model, enabling it to adapt to the task of video frame interpolation. Benefiting from the base model trained on a large-scale video dataset, our video interpolation model is capable of achieving reasonable interpolation results even in cases of large motion. This makes our approach more suitable for real-world scenarios.

References

[1] 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

[2] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023

[3] T. Xue, B. Chen, J. Wu, D. Wei, W. T. Freeman“, Video Enhancement with Task-Oriented Flow,”IJCV 2019. International Journal of Computer Vision, vol 127. 2019, pp 1106-1125.

[4] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, L. V. Gool, “ The 2017 DAVIS Challenge on Video Object Segmentation,” arXiv preprint, arXiv : 1704.00675.

[5] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In Proceedings of the European Conference on Computer Vision (ECCV), 2022.

[6] 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 ECCV, 2022.

[7] Danier, Duolikun, Fan Zhang, and David Bull. ”Ldmvfi: Video frame interpolation with latent diffusion models.” Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 2. 2024.

[8] Jain S, Watson D, Tabellion E, et al. Video interpolation with diffusion models Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 7341-7351.

School of Fundamental Science and Engineering
Waseda University, Shillman Hall 401, 3-14-9 Okubo
Shinjuku Tokyo, 169-0072 Phone: 03-5286-2509
Fax:03-5286-3488