VFI-LoRA : Leveraging Video Diffusion Models for Video Interpolation Through LoRA Finetuning

1st Taiju Watanabe
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan

2st Takahiro Shindo
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan

3st Yui Tatsumi
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan

4st Hiroshi Watanabe
Graduate School of Fundamental Science and Engineering
Waseda University
Tokyo, Japan

***Abstract*—We present VFI-LoRA, a method for generating precise and realistic intermediate frames given only a first and a last frame. By leveraging a pretrained image-to-video diffusion model (Stable Video Diffusion), our approach produces frames that remain both spatially and temporally consistent. To adapt the diffusion model for video frame interpolation, we employ Low-Rank Adaptation (LoRA) to efficiently finetune the model, enabling it to handle large motions effectively. Additionally, we introduce a technique to refine generated frames during the denoising stages of inference. To address scenarios where objects may disappear in sequences with large motions, we further perform renosing and denosing steps after generating latent features with backward process of inference. We compare our method against both existing diffusion-based and CNN-based interpolation methods, demonstrating its effectiveness, particularly for sequences with complex, nonlinear motions.**

***Index Terms*—Video frame interpolation, Stable Video Diffusion, LoRA**

I. Introduction

Video frame interpolation is one of the core task in computer vision, aiming to generate intermediate frames using only the first and last frames [1]. This technology has a broad range of real-world applications, including frame-rate enhancement and video compression. Advances in deep neural networks have fueled significant progress in this field, with existing approaches primarily relying on convolutional neural networks (CNNs) [2]–[5]. Meanwhile, recent advancements in large-scale diffusion models, such as text-to-video [6] and image-to-video [7] models, have demonstrated the capability to generate videos with significant motion. Compared to developing diffusion-based interpolation models from scratch, leveraging these pretrained models allows for training with smaller datasets and reduces computational complexity. To harness the capabilities of pretrained diffusion models, we propose VFI-LoRA, a method for video frame interpolation that builds upon a pretrained image-to-video diffusion model. Our approach leverages the pretrained Stable Video Diffusion (SVD) model [7], finetuning it with Low-Rank Adaptation (LoRA) [8] to preserve its knowledge while generating high-quality frame predictions. Moreover, to address object disappearance in sequences with large motions, we introduce Frame Refinement. This strategy injects noise into features generated by the 3D U-Net after the backward inference process and then re-denoises them, producing finer details that better approximate the true data distribution. Extensive experiments demonstrate that our LoRA-based finetuning and Frame Refinement outperform existing diffusion-based and CNN-based methods.

These research results were obtained from the commissioned research (JPJ012368C05101) by National Institute of Information and Communications Technology (NICT), JAPAN.

II. Related work

A. Video Frame Interpolation (VFI)

Video frame interpolation (VFI), the process of generating intermediate frames between existing frames of a video sequence, has become a critical technique in the field of computer vision [1]. By artificially increasing the frame rate, frame interpolation enhances the smoothness, making it especially valuable for a wide range of applications such as slow-motion playback, frame rate upconversion and video compression. Breakthroughs in deep neural networks have led certain interpolation approaches to incorporate CNNs within their frameworks. Such techniques typically involve using CNNs to estimate motion vectors between frames. These estimated vectors are then employed to warp existing frames, enabling the generation of intermediate frames based on the computed motion. However, CNN-based techniques are often inadequate for processing sequences that involve large motions. To effectively handle sequences with large motions, recent studies have introduced diffusion-based methods.

B. Latent Diffusion Models (LDMs)

Latent Diffusion Models (LDMs) [9] perform diffusion steps within a lower-dimensional latent space learned by an autoencoder. An autoencoder is first trained so that its encoder $\mathcal{E}$ maps an input image $\mathbf{x} \in \mathbb{R}^D$ to a latent representation $\mathbf{z} \in \mathbb{R}^d$ (with $d \ll D$), and its decoder $\mathcal{D}$ reconstructs $\mathbf{x}$

Fig. 1. Forward and backward predictions from SVDKFI.

from $\mathbf{z}$. Once the autoencoder is trained, the diffusion process proceeds on the latent representation $\mathbf{z}$. At each diffusion step t, Gaussian noise is added according to

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}\big(\mathbf{z}_t \mid \sqrt{1-\beta_t}\,\mathbf{z}_{t-1}, \beta_t \mathbf{I}\big). \tag{1}$$

Here, β_t is the variance schedule parameter at step t. It controls the amount of noise introduced into the latent representation $\mathbf{z}_{t-1}$ to obtain $\mathbf{z}_t$. A corresponding reverse process $p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t)$ is also assumed to be Gaussian, and it is learned via a neural network $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)$ that predicts the added noise. The training objective for this noise-predicting network typically takes the form

$$L(\theta) = \mathbb{E}_{\mathbf{z}_0,\, \boldsymbol{\epsilon},\, t}\big[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t; t)\|^2\big], \tag{2}$$

where $\boldsymbol{\epsilon}$ is the actual noise at step t, and $\mathbf{z}_t$ is the noisy version of the latent representation. Sampling from an LDM involves first drawing a latent $\mathbf{z}_T$ from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Then, using the learned reverse process, one iteratively removes noise step by step to generate $\mathbf{z}_0$. Finally, this denoised latent vector is passed through the decoder $\mathcal{D}$ to produce the resulting image $\mathbf{x}_{\text{sample}}$ in the pixel space.

C. Diffusion-based Video Interpolation

Multiple studies have exploited diffusion-based methods to VFI, leveraging the capacity to produce visually appealing intermediate frames [10], [11]. Recent advances in image-to-video diffusion models, such as Stable Video Diffusion (SVD), have introduced sophisticated sampling strategies that integrate both forward and backward predictions. For instance, TRF [12] integrates forward and backward conditioned denoising without finetuning SVD. Similarly, SVDKFI [13] rotates attention maps of SVD and finetunes the temporal layers conditioned on the last frame to generate a model for backward prediction. During inference, both the forward and backward models generate noise, which is then merged to produce time-consistent features. However, we observed that backward predictions are less effective at generating synthetic frames than the standard forward prediction model. Fig. 1 compares outputs from both the forward model and the backward model of SVDKFI. As illustrated in Fig. 1, the object (car) vanishes in the frames produced by the backward prediction. We hypothesize that this issue arises because the pretrained SVD model was primarily trained for forward prediction. Finetuning this model for backward prediction may disrupt the temporal knowledge necessary to produce temporally consistent frames.

D. Low-Rank Adaptation (LoRA)

Generative AI, including text-to-video diffusion models [6], and image-to-video diffusion models [7] have gained prominence. Alongside these advancements, efficient finetuning techniques have been developed. Low-Rank Adaptation (LoRA) [8] serves as an example of an efficient finetuning technique. LoRA focuses on learning the differences between the original model weights and the finetuned weights. Finetuning adjusts model parameters from their original values θ to new values θ' by learning a parameter update $\Delta\theta$, $\theta' = \theta + \Delta\theta$. Given that the original parameters θ reside in $\mathbb{R}^{d\times k}$, the approach factorizes the update $\Delta\theta$ as $\Delta\theta = BA$, where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$(with $r \ll d$). This low-rank factorization significantly reduces the number of parameters that need to be learned during finetuning.

III. Proposed Method

A. LoRA Finetuning for Video Interpolation

We propose a method, VFI-LoRA, for leveraging a pretrained image-to-video diffusion model, Stable Video Diffusion (SVD) for video interpolation using Low-Rank Adaptation (LoRA). As shown in Fig. 2, our finetuning phase diverges from previous work like SVDKFI, which requires finetuning for backward prediction. Instead, we finetune SVD to generate videos that progress forward in time while conditioning on the last frame. This approach preserves the model's learned forward motion dynamics, which were acquired from abundant video data. Given a first frame I_0 and a last frame I_{N-1}, our objective is to generate N frames: $I_0, I_1, \ldots, I_{N-1}$. During finetuning, we concatenate these frames to form an input tensor $\mathbf{x} \in \mathbb{R}^D$. The encoder $\mathcal{E}$ of a pretrained Variational Autoencoder (VAE) [14] maps the input tensor $\mathbf{x}$ to a latent representation $\mathbf{z} \in \mathbb{R}^d (d \ll D)$. To simulate the diffusion process, Gaussian noise is added to the latent representation at time step t, resulting in

$$\mathbf{z}_t = \alpha_t \mathbf{z} + \sigma_t \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is Gaussian noise, and α_t and σ_t are scalars defined by a noise schedule. Additionally, the first and last frames, I_0 and I_{N-1}, are encoded into embeddings using a pretrained CLIP Image Encoder ($\mathcal{E}_{CLIP}$) [15]. These embeddings, $c_0, c_{N-1} \in \mathbb{R}^{d'} (d' \ll D)$, serve as conditions for the 3D U-Net $\boldsymbol{\epsilon}_\theta$ within the SVD model. We incorporate LoRA into the 3D U-Net, which dramatically reduces the number of

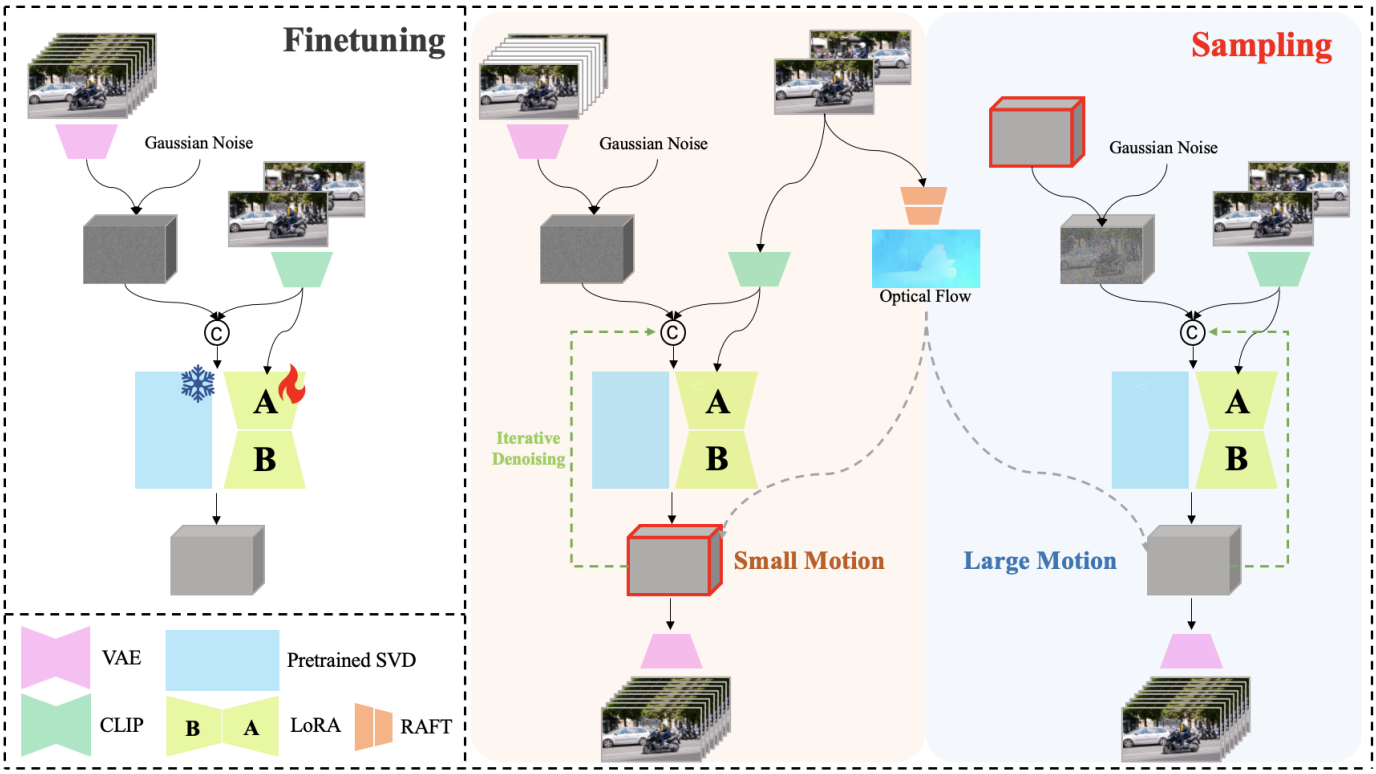

Fig. 2. Overview of our VFI-LoRA.

trainable parameters. The network is optimized by minimizing the loss function,

$$L(\theta) = \mathbb{E}_{\mathbf{z}_0,\,\boldsymbol{\epsilon},\,t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t; t, c_0, c_{N-1})\|^2\right]. \tag{4}$$

B. Frame Refinement (FR)

During inference, the latent representation $\mathbf{z}$ input to the 3D U-Net with LoRA is initialized by random Gaussian noise. 3D U-Net then receives embeddings $c_0, c_{N-1} \in \mathbb{R}^{d'}$ from the CLIP Image Encoder ($\mathcal{E}_{CLIP}$), which serve as conditions. As shown in Fig. 4, our experiments reveal that for sequences with extreme motions, the proposed method struggles to generate objects faithfully. To address this, we introduce Frame Refinement. For sequences exhibiting significant motion, after completing the standard diffusion process, an additional diffusion process is executed to correct occlusions or artifacts that occurred during the initial denoising steps. The detailed procedure is outlined in Fig. 2 and Alg. 1. The Frame Refinement process begins by estimating the optical flow between the first and last frames using RAFT ($\mathcal{F}$) [16]. Sequences are subsequently ranked based on the magnitude of motion. Sequences exhibiting motion magnitudes above the threshold τ are classified as large-motion sequences, while others are categorized as small-motion sequences. For large-motion sequences, the latent representation is re-noised for a few steps by reintroducing Gaussian noise according to the scheduler. Then, iterative denoising is performed again to yield the final latent features. This method aims to guide latent features that diverged from the true distribution back towards a more accurate representation, using insights gained from the initial diffusion stage. Finally, the refined latent representation is fed into the decoder $\mathcal{D}$ of a pretrained VAE to generate the final output frames. This Frame Refinement step enhances the quality of generated frames, particularly for sequences with large motions.

IV. Experiment

A. Implementation Details

We trained our method using the high-quality OpenVid-1M dataset [17] with an image resolution of 512×512 pixels. Specifically, our 3D U-Net is trained to generate all 9 frames given only the first and last frames. For finetuning the 3D U-Net, we apply LoRA with a rank of 8 to adapt the weights of linear and convolutional layers. This approach dramatically reduces the number of trainable parameters from 1.52B to 14M. The training procedure runs for 30,000 iterations using the AdamW optimizer [18] with a learning rate of 1e-4. We use a batch size of 2 and apply gradient accumulation over 4 steps to stabilize training. During inference, we first use RAFT to extract the optical flow between the first and last frames. Sequences are then ranked by the magnitude of motion, with the top 10% of sequences classified as large-motion and the remainder as small-motion sequences. For all sequences, an iterative denoising process is performed for 50 timesteps using the Euler Discrete Scheduler [19]. For sequences categorized as having large motions, an additional re-noising step is applied for 35 timesteps after the diffusion process, followed

TABLE I
QUANTITATIVE COMPARISON ON DAVIS AND VISDRONE-VID DATASETS.

	DAVIS					VisDrone-VID				
Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
RIFE	19.87	0.6571	0.2869	14.90	391.4	26.34	0.8557	0.1338	7.338	140.9
FILM	19.68	0.6458	**0.2764**	**11.53**	341.1	26.54	0.8568	**0.1253**	**5.935**	96.17
AMT	**20.16**	**0.6729**	0.3160	25.06	391.5	**27.05**	**0.8662**	0.1424	10.77	106.9
LDMVFI	19.30	0.6244	0.3173	17.97	460.0	24.78	0.8265	0.1554	9.489	194.8
TRF	15.72	0.5146	0.4495	32.80	570.9	17.39	0.5886	0.3626	37.75	645.0
SVDKFI	16.35	0.5265	0.3781	27.14	300.3	19.32	0.6478	0.2604	31.33	297.8
VFI-LoRA w/o FR	17.64	0.5796	0.3272	24.37	155.1	21.25	0.7343	0.2107	26.27	68.50
VFI-LoRA w/ FR	17.56	0.5768	0.3305	23.60	**150.5**	21.29	0.7348	0.2111	25.77	**66.34**

Algorithm 1 Diffusion sampling using Frame Refinement.

Require: $I_0, I_{N-1}, \boldsymbol{\epsilon}_\theta, \mathcal{E}_{CLIP}(\cdot), \mathcal{D}(\cdot), \mathcal{F}$
Define optical flow threshold τ using $\mathcal{F}$
Generate c_0, c_{N-1} from $\mathcal{E}_{CLIP}(I_0), \mathcal{E}_{CLIP}(I_{N-1})$
Set $z_T \sim \mathcal{N}(0, I)$
Denosing steps:
for $t \leftarrow T$ **to** 1 **do**
  $\hat{v}_t = \boldsymbol{\epsilon}_\theta(z_t; t, c_0, c_{N-1})$
  $z_{t-1} = \text{update}(z_t, \hat{v}_t; t)$
end for
if $\mathcal{F}(I_0, I_{N-1}) > \tau$ **then**
  Perform K steps of renosing:
  for $k \leftarrow 1$ **to** K **do**
    $\hat{v}'_{k-1} = \boldsymbol{\epsilon}_\theta(z_{k-1}; k-1, c_0, c_{N-1})$
    $z_k = \text{inverse\_update}(z_{k-1}, \hat{v}'_{k-1}; k-1)$
  end for
  Additional denosing steps:
  for $k \leftarrow K$ **to** 1 **do**
    $\hat{v}'_k = \boldsymbol{\epsilon}_\theta(z_k; k, c_0, c_{N-1})$
    $z_{k-1} = \text{update}(z_k, \hat{v}'_k; k)$
  end for
end if
Return $\mathcal{D}(z_0)$

by further iterative denoising. This additional step aims to refine frames affected by substantial motion artifacts.

B. Evaluation

For evaluation, we use the DAVIS [20] and VisDrone-VID [21] datasets with a resolution of 448×832 pixels. Unlike previous works that cropped DAVIS sequences to 256×256 (which is unsuitable for the original SVD model trained on high-resolution data), we retain the high-resolution sequences. Given the first and last images of a sequence, models perform multiple frame interpolations to generate 7 intermediate frames, and performance is evaluated based on these generated frames. We compare our method with the CNN-based methods, RIFE [2], FILM [3], AMT [4] and diffusion-based methods, LDMVFI [10], TRF [12], SVDKFI [13].

TABLE II
IMPACT OF FRAME REFINEMENT ON SVDKFI.

	DAVIS				
Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
SVDKFI	16.35	0.5265	0.3781	27.14	300.3
SVDKFI w/ FR	**16.41**	**0.5298**	**0.3766**	**26.74**	**289.8**

C. Qualitative Comparison

Qualitative comparisons of intermediate frames generated by our method alongside those produced by previous approaches is presented in Fig. 3. Our VFI-LoRA produces frames that are more detailed and temporally smooth, offering a more coherent result. CNN-based approaches (FILM, AMT) fail to maintain spatial and temporal coherence, resulting in occlusions. While SVDKFI captures some temporal information, it struggles with accurately modeling complex motions, leading to distortions in the generated frames. In contrast, our method especially with FR preserves the shape of objects, achieving both spatial and temporal consistency. Fig. 4 highlights the benefits of our Frame Refinement approach in scenarios with extreme motions. Without FR, the object (person) disappears from the generated sequence. However, with Frame Refinement, VFI-LoRA successfully generates the person while accurately capturing the correct motion, thereby demonstrating its effectiveness in handling large-motion sequences.

D. Quantitative Comparison

For quantitative evaluation, we employed several metrics, PSNR, SSIM [22], LPIPS [23], FID [24], and FVD [25] to assess the quality of 7 generated frames. Table 1 presents the results on the DAVIS and VisDrone-VID datasets. While CNN-based methods such as FILM and AMT achieve the highest scores in PSNR, SSIM, LPIPS, and FID, our method excels in terms of FVD on both datasets. As noted in VIDIM [11], metrics like PSNR, SSIM, LPIPS, and FID may not fully capture the perceptual quality of video interpolation results, as they fail to consider temporal consistency across frames. Furthermore, our proposed FR approach results in a modest improvement in FVD, as it is applied exclusively to sequences with large motions.

Fig. 3. Qualitative comparison on intermediate frames.

Fig. 4. Comparison of Frame Refinement (FR) in sequence with extreme motion.

E. Ablation Study on Frame Refinement

We conducted an ablation study on Frame Refinement by integrating it with SVDKFI. Table 2 presents the quantitative comparison of SVDKFI with and without Frame Refinement. The results in Table 2 indicate that our proposed sampling scheme improves SVDKFI across all evaluated metrics. This enhancement suggests that our Frame Refinement technique is not only beneficial for VFI-LoRA but can also be effectively applied to other diffusion-based interpolation methods.

V. Conclusion

In this paper, we present VFI-LoRA, a method for generating visually faithful intermediate frames given only the first and last frames of a video. By leveraging Stable Video Diffusion (SVD) with Low-Rank Adaptation (LoRA), our approach produces frames that are both spatially and temporally consistent. Furthermore, we introduce Frame Refinement, which improves the quality of generated frames during the denoising stages of inference, particularly in sequences with large motions. Through extensive experiments, we demonstrate that VFI-LoRA outperforms existing methods in terms of frame quality. Additionally, Frame Refinement can be adapted to other diffusion-based interpolation methods, underscoring its potential applicability in future approaches.

References

[1] J. Dong, K. Ota and M. Dong, "Video Frame Interpolation: A Comprehensive Survey," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 19, pp. 1-31, 2023.

[2] Z. Huang, T. Zhang, W. Heng, B. Shi and S. Zhou, "Real-Time Intermediate Flow Estimation for Video Frame Interpolation," European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 13674, pp. 624-642, 2022.

[3] F. Reda, J. Kontkanen, E. Tabellion, D. Sun, C. Pantofaru and B. Curless, "FILM: Frame Interpolation for Large Motion," European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 13667, pp. 250-266, 2022.

[4] Z. Li, Z. -L. Zhu, L. -H. Han, Q. Hou, C. -L. Guo and M. -M. Cheng, "AMT: All-Pairs Multi-Field Transforms for Efficient Frame Interpolation," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9801-9810, 2023.

[5] Z. Shi, X. Liu, K. Shi, L. Dai and J. Chen, "Video Frame Interpolation via Generalized Deformable Convolution," in IEEE Transactions on Multimedia, vol. 24, pp. 426-439, 2022.

[6] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, G. Liu, A. Raj, Y. Li, M. Rubinstein, T. Michaeli, O. Wang, D. Sun, T. Dekel and I. Mosseri, "Lumiere: A Space-Time Diffusion Model for Video Generation," SA '24: SIGGRAPH Asia 2024 Conference Papers, pp. 1-11, 2024.

[7] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani and R. Rombach, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets," arXiv preprint arXiv:2311.15127, 2023.

[8] E. J. Hu, y. shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," International Conference on Learning Representations (ICLR), 2022.

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10674-10685, 2022.

[10] D. Danier, F, Zhang, and David Bull, "LDMVFI: Video frame interpolation with latent diffusion models," In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 1472–1480, 2024.

[11] S. Jain, D. Watson, E. Tabellion, A. Hołyński, B. Poole and J. Kontkanen, "Video Interpolation with Diffusion Models," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7341-7351, 2023.

[12] H. Feng, Z. Ding, Z. Xia, S. Niklaus, V. Abrevaya, M. J. Black and X. Zhang, "Explorative Inbetweening of Time and Space," European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol 15136, pp. 378–395, 2024.

[13] X. Wang, B. Zhou, B. Curless, I. Kemelmacher-Shlizerman, A. Holynski and S. M. Seitz, "Generative Inbetweening: Adapting Image-to-Video Models for Keyframe Interpolation," arXiv preprint arXiv:2408.15239, 2024.

[14] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," International Conference on Learning Representations (ICLR), 2014.

[15] A. Radford, J. Wook Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," arXiv preprint arXiv:2103.00020, 2021.

[16] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol 12347, pp. 402-419, 2020.

[17] K. Nan, R. Xie, P. Zhou, T. Fan, Z. Yang, Z. Chen, X. Li, J. Yang and Y. Tai, "OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-video Generation," arXiv preprint arXiv:2407.02371, 2024.

[18] I. Loshchilov, F. Hutter, "Decoupled Weight Decay Regularization," International Conference on Learning Representations (ICLR), 2019.

[19] T. Karras, M. Aittala, S. Laine, T. Aila, "Elucidating the design space of diffusion-based generative models," Proceedings of the 36th International Conference on Neural Information Processing Systems, pp. 26565-26577, 2022.

[20] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung and L. V. Gool, "The 2017 DAVIS Challenge on Video Object Segmentation," arXiv preprint arXiv:1704.00675, 2024.

[21] P. Zhu, D. Du, L. Wen, X. Bian, H. Ling, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang, L. Bo, H. Shi, R. Zhu, B. Dong, D. Pailla, F. Ni, G. Gao, G. Liu, H. Xiong, J. Ge, J. Zhou, J. Hu, L. Sun, L. Chen, M. Lauer, Q. Liu, S. Chennamsetty, T. Sun, T. Wu, V. Kollerathu, W. Tian, W. Qin, X. Chen, X. Zhao, Y. Lian, Y. Wu, Y. Li, Y. Li, Y. Wang, Y. Song, Y. Yao, Y. Zhang, Z. Pi, Z. Chen, Z. Xu, Z. Xiao, Z. Luo and Z. Liu., "VisDrone-VID2019: The Vision Meets Drone Object Detection in Video Challenge Results," IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 227-235, 2019.

[22] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," in IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.

[23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 586-595, 2018.

[24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6629-6640, 2017.

[25] S. Ge, A. Mahapatra, G. Parmar, J. -Y. Zhu and J. -B. Huang, "On the Content Bias in Fréchet Video Distance," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7277-7288, 2024.