Controllable Tracking-Based Video Frame Interpolation (Supplementary Document)

KARLIS MARTINS BRIEDIS, DisneyResearch|Studios, Switzerland and ETH Zürich, Switzerland
ABDELAZIZ DJELOUAH, DisneyResearch|Studios, Switzerland
RAPHAËL ORTIZ, DisneyResearch|Studios, Switzerland
MARKUS GROSS, DisneyResearch|Studios, Switzerland and ETH Zürich, Switzerland
CHRISTOPHER SCHROERS, DisneyResearch|Studios, Switzerland

A Additional Results

To evaluate multi-frame interpolation quality, we follow Jain et al. [2024] and generate sequences of 9 frames from the DAVIS [Perazzi et al. 2016] dataset, and report the mean over all 7 interpolated intermediate frames. However, we use the higher-resolution *1080p* images and evaluate on full images at their original resolution. Results are reported in Table 1.

Additional quantitative evaluation on SNU-FILM [Choi et al. 2020], X-TEST [Sim et al. 2021], and Vimeo-90K-septuplet [Xue et al. 2019] datasets is provided in Tables 2 and 3.

Finally, in Figure 2 we show a visualization of PSNR and LPIPS tradeoff of different methods over the DAVIS dataset, and in Figure 3 provide a qualitative comparison between different sharpness control values.

B User Study

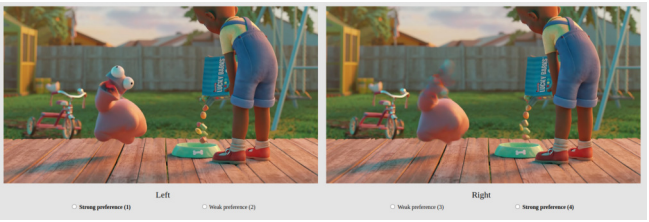

Fig. 1. A screenshot of the interface used in our user study. The user is prompted to select if they prefer the left or right video, looping back and forth, and indicate if it is a strong or weak preference. Images contain assets from The Daily Dweebs by Blender Foundation.

To evaluate the perceptual improvement of our assisted and non-assisted versions, we conduct a web-based user study.

Study Interface. In this study, participants were asked to "select which result *they* think is better, e.g. it looks more natural, has fewer artifacts, etc., and indicate if it is a strong or weak preference" as shown in Figure 1. Each participant was asked to do at minimum 40 comparisons and optionally continue to do all comparisons. Each sequence, method pair, and the order it appears on the screen was sampled at random.

Method Selection. For the comparisons, we select the top performing methods in our quantitative evaluation based on PSNR and LPIPS scores - GIMM, CFA-RIFE, PerVFI, FILM-$\mathcal{L}_S$ -, as well as a generative method - LDMVFI. We compare every method with our unassisted and our assisted interpolation results, as well as compare them with each other.

Data Selection. To select a fair set of sequences for all methods, we choose to sample 10 sequences from the DAVIS dataset, using first frames as in our quantitative evaluation.

However, for several sequences all methods show a very good quality reconstruction, thus, to avoid comparing almost equal images, we decided to **not** sample from the following sequences - `bear`, `blackswan`, `boat`, `bus`, `car-turn`, `dog`, `elephant`, `goat`, `hike`,

Table 1. Quantitative evaluation of multi-frame interpolation without using any user inputs.

	DAVIS-7 1080p		
	PSNR ↑	SSIM ↑	LPIPS ↓
SoftSplat-$\mathcal{L}_1$ [Niklaus and Liu 2020]	19.02	0.539	0.3461
XVFI-Vimeo [Sim et al. 2021]	18.77	0.544	0.4555
ABME [Park et al. 2021]	19.42	0.579	0.4566
VFIFormer [Lu et al. 2022]	Out of Memory		
RIFE [Huang et al. 2022]	19.22	0.540	0.2986
FILM-L_1 [Reda et al. 2022]	19.19	0.548	0.3228
AMT-G [Li et al. 2023]	19.27	0.573	0.4114
UPRNet LARGE [Jin et al. 2023]	18.93	0.552	0.3983
EMA-VFI [Zhang et al. 2023]	19.57	0.574	0.4609
SGM 50% [Liu et al. 2024]	17.75	0.485	0.4270
CFA-RIFE [Zhong et al. 2025]	20.00	0.593	0.4223
VFIMamba [Zhang et al. 2024]	19.86	**0.595**	0.4659
GIMM [Guo et al. 2024]	**20.82**	0.589	**0.2232**
Ours-$\mathcal{S}_{0.0}$	20.00	0.572	0.3009
SoftSplat-$\mathcal{L}_F$ [Niklaus and Liu 2020]	18.71	0.503	0.3049
FILM-$\mathcal{L}_S$ [Reda et al. 2022]	18.99	0.525	0.2887
PerVFI [Wu et al. 2024]	19.66	0.540	0.2626
LDMVFI [Danier et al. 2024]	18.78	0.513	0.3191
Ours-$\mathcal{S}_{1.0}$	**19.86**	**0.550**	**0.2495**

SIGGRAPH Conference Papers '25, Vancouver, BC, Canada

ACM ISBN 979-8-4007-1540-2/2025/08
https://doi.org/10.1145/3721238.3730598

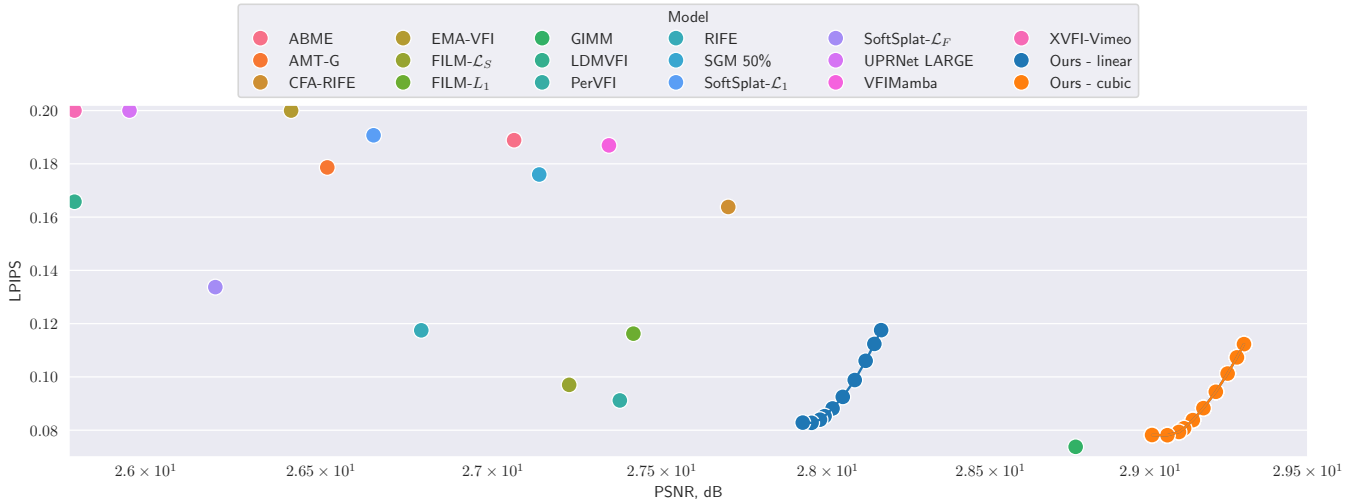

Fig. 2. PSNR (in logarithmic scale, clipped to a minimum of 27 dB) and LPIPS (clipped to a maximum of 0.18) values for different methods and different perceptual control values of our method, evaluated on the DAVIS test dataset.

Fig. 3. Qualitative results for different sharpness control values $\mathcal{S}_w$. With increased parameter values, outputs have higher level of detail, however, it is not fully matching the reference. Image from the DAVIS dataset.

`kite-walk`, `mallard-fly`, `mallard-water`, `motocross-bumps`, `motorbike`, `paragliding`, `paragliding-launch`, `rhino`, `scooter-black`, `soapbox`, and `tennis`.

Finally, we sample the following 10 scenes: `breakdance`, `dog-agility`, `drift-chicane`, `drift-turn`, `horsejump-low`, `parkour`, `scooter-gray`, `stroller`, `surf`, `swing`. We choose to add a frame triplet from animated movie The Daily Dweebs as 11-th comparison to represent a different content type with very non-linear movement.

Assisted Interpolation. We then run our user interaction tool to obtain assisted outputs of our method. For several sequences in the sampled set of comparisons, our baseline version already shows very good results and no interaction is necessary. The following samples are not included for comparisons with our assisted version - `drift-chicane`, `parkour`, `stroller`, `surf`.

Results. In total, 26 participants cast 1598 votes, with a minimum of 12 different users voting for each distinct query. The full result breakdown per sequence and comparing method pair is provided in Table 4. The interaction time includes the interpolation time for all intermediate low-framerate previews but excludes the time for the final 32× rendering.

References

Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. 2020. Channel Attention Is All You Need for Video Frame Interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, Number 07. 10663–10671.

Duolikun Danier, Fan Zhang, and David Bull. 2024. LDMVFI: Video Frame Interpolation with Latent Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 2 (March 2024), 1472–1480. https://doi.org/10.1609/aaai.v38i2.27912

Zujin Guo, Wei Li, and Chen Change Loy. 2024. Generalizable Implicit Motion Modeling for Video Frame Interpolation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems.*

Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. 2022. Real-Time Intermediate Flow Estimation for Video Frame Interpolation. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Vol. 13674. Springer Nature Switzerland, Cham, 624–642. https://doi.org/10.1007/978-3-031-19781-9_36

Siddhant Jain, Daniel Watson, Eric Tabellion, Aleksander Holynski, Ben Poole, and Janne Kontkanen. 2024. Video Interpolation with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 7341–7351.

Xin Jin, Longhai Wu, Jie Chen, Youxin Chen, Jayoon Koo, and Cheul-hee Hahm. 2023. A Unified Pyramid Recurrent Network for Video Frame Interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 1578–1587.

Table 2. Quantitative evaluation against prior methods on SNU-FILM [Choi et al. 2020] datasets.

	SNU-Easy			SNU-Medium			SNU-Hard			SNU-Extreme		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
SoftSplat-$\mathcal{L}_1$ [Niklaus and Liu 2020]	40.24	**0.984**	0.0185	36.06	0.966	0.0335	30.50	0.900	0.0635	25.14	0.787	0.1308
XVFI-Vimeo [Sim et al. 2021]	40.00	0.983	0.0177	35.37	0.963	0.0322	29.56	0.883	0.0751	24.14	0.765	0.1550
ABME [Park et al. 2021]	39.74	0.983	0.0228	35.85	0.966	0.0380	30.62	0.901	0.0668	25.44	0.792	0.1271
VFIFormer [Lu et al. 2022]	40.28	**0.984**	0.0180	36.08	**0.967**	0.0337	30.28	0.898	0.0691	24.96	0.786	0.1461
RIFE [Huang et al. 2022]	39.74	0.982	0.0131	35.45	0.962	0.0236	29.93	0.891	0.0481	24.86	0.777	0.0981
FILM-L_1 [Reda et al. 2022]	40.19	**0.984**	0.0186	36.03	0.966	0.0320	30.49	0.899	0.0575	25.20	0.785	0.1068
AMT-G [Li et al. 2023]	40.10	**0.984**	0.0198	35.91	0.966	0.0331	30.42	0.899	0.0606	25.06	0.786	0.1214
UPRNet LARGE [Jin et al. 2023]	40.33	**0.984**	0.0188	36.19	**0.967**	0.0343	30.50	0.900	0.0672	24.99	0.785	0.1433
EMA-VFI [Zhang et al. 2023]	40.19	**0.984**	0.0185	36.14	**0.967**	0.0335	30.65	0.899	0.0664	25.27	0.785	0.1402
SGM 50% [Liu et al. 2024]	40.36	**0.984**	0.0186	36.13	0.966	0.0326	30.64	0.899	0.0633	25.38	0.788	0.1223
CFA-RIFE [Zhong et al. 2025]	40.09	**0.984**	0.0190	35.93	0.965	0.0325	30.47	0.899	0.0614	25.42	0.790	0.1236
VFIMamba [Zhang et al. 2024]	**40.44**	**0.984**	0.0184	**36.23**	**0.967**	0.0338	30.74	0.902	0.0651	25.51	**0.794**	0.1267
GIMM [Guo et al. 2024]	40.13	0.983	**0.0105**	36.09	0.966	**0.0188**	**30.86**	**0.903**	**0.0382**	**25.71**	0.792	**0.0779**
Ours-$\mathcal{S}_{0.0}$	39.63	0.983	0.0190	36.00	0.966	0.0328	30.84	0.902	0.0591	25.63	0.792	0.1112
SoftSplat-$\mathcal{L}_F$ [Niklaus and Liu 2020]	39.90	0.982	**0.0109**	35.71	0.963	**0.0199**	30.19	0.892	**0.0425**	24.80	0.770	0.0973
FILM-$\mathcal{L}_S$ [Reda et al. 2022]	**40.14**	**0.983**	0.0120	**35.91**	**0.965**	0.0215	30.37	0.895	0.0432	25.09	0.778	0.0891
PerVFI [Wu et al. 2024]	38.07	0.974	0.0141	34.59	0.955	0.0245	29.82	0.887	0.0470	25.03	0.775	0.0902
LDMVFI [Danier et al. 2024]	38.74	0.979	0.0145	34.04	0.950	0.0284	28.57	0.868	0.0599	23.94	0.751	0.1224
Ours-$\mathcal{S}_{1.0}$	39.47	0.982	0.0128	35.80	**0.965**	0.0221	**30.63**	**0.898**	0.0432	**25.44**	**0.785**	**0.0857**

Table 3. Quantitative evaluation against prior methods on Vimeo-90K-7f [Xue et al. 2019] and X-TEST [Sim et al. 2021] datasets.

	Vimeo-90K-7f			X-TEST-2K			X-TEST-4K		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
SoftSplat-$\mathcal{L}_1$ [Niklaus and Liu 2020]	35.75	0.958	0.0312	28.97	0.807	0.1456	24.81	0.736	0.2970
XVFI-Vimeo [Sim et al. 2021]	34.79	0.951	0.0313	25.17	0.692	0.2483	22.77	0.690	0.3012
ABME [Park et al. 2021]	35.84	0.958	0.0309	29.15	0.813	0.1613	Out of Memory		
VFIFormer [Lu et al. 2022]	36.19	**0.960**	0.0297	Out of Memory			Out of Memory		
RIFE [Huang et al. 2022]	34.04	0.945	0.0233	28.95	0.803	0.0882	25.36	0.716	0.1989
FILM-L_1 [Reda et al. 2022]	35.83	0.958	0.0278	30.33	0.838	0.0772	Out of Memory		
AMT-G [Li et al. 2023]	36.16	**0.960**	0.0279	29.26	0.804	0.1540	Out of Memory		
UPRNet LARGE [Jin et al. 2023]	36.11	0.959	0.0292	27.12	0.752	0.2389	Out of Memory		
EMA-VFI [Zhang et al. 2023]	**36.23**	0.959	0.0292	28.11	0.754	0.2193	Out of Memory		
SGM 50% [Liu et al. 2024]	35.54	0.956	0.0297	29.34	0.798	0.1530	Out of Memory		
CFA-RIFE [Zhong et al. 2025]	34.81	0.952	0.0307	30.43	0.840	0.1015	27.31	0.772	0.2377
VFIMamba [Zhang et al. 2024]	**36.23**	0.959	0.0289	31.03	0.855	0.1011	Out of Memory		
GIMM [Guo et al. 2024]	36.14	**0.960**	**0.0150**	**31.69**	**0.860**	**0.0524**	**30.87**	**0.839**	**0.1116**
Ours-$\mathcal{S}_{0.0}$	35.49	0.956	0.0296	30.28	0.847	0.0667	29.34	0.813	0.1166
SoftSplat-$\mathcal{L}_F$ [Niklaus and Liu 2020]	35.17	0.950	**0.0178**	28.35	0.781	0.0925	24.59	0.697	0.2075
FILM-$\mathcal{L}_S$ [Reda et al. 2022]	**35.64**	**0.955**	0.0183	**30.28**	0.835	0.0546	Out of Memory		
PerVFI [Wu et al. 2024]	33.54	0.939	0.0241	29.82	0.831	0.0489	Out of Memory		
LDMVFI [Danier et al. 2024]	33.39	0.938	0.0258	23.92	0.642	0.1915	Out of Memory		
Ours-$\mathcal{S}_{1.0}$	35.25	0.953	0.0192	30.11	**0.842**	**0.0449**	**29.12**	**0.804**	**0.0757**

Table 4. Full user study results. For each sequence the first row corresponds to our unassisted method and the second row corresponds to the user-assisted output. In each cell, we show the percentage of votes as (strong preference for ours | weak preference for ours | strong preference for theirs | weak preference for theirs).

Sequence	FILM	LDMVFI	PerVFI	CFA-RIFE	GIMM	Ours (unassisted)	Interaction Time
breakdance	14 | 57 | 29 | 0	64 | 36 | 0 | 0	31 | 62 | 8 | 0	0 | 31 | 54 | 15	0 | 8 | 77 | 15	–	
	71 | 21 | 7 | 0	77 | 23 | 0 | 0	75 | 25 | 0 | 0	31 | 54 | 15 | 0	23 | 62 | 15 | 0	38 | 62 | 0 | 0	06:59
dog-agility	0 | 77 | 23 | 0	38 | 62 | 0 | 0	38 | 54 | 8 | 0	38 | 62 | 0 | 0	0 | 23 | 54 | 23	–	
	0 | 83 | 17 | 0	50 | 43 | 7 | 0	47 | 47 | 7 | 0	69 | 31 | 0 | 0	8 | 69 | 23 | 0	15 | 62 | 23 | 0	07:52
drift-chicane	17 | 59 | 24 | 0	93 | 7 | 0 | 0	61 | 39 | 0 | 0	24 | 69 | 7 | 0	7 | 55 | 38 | 0	–	
	-	-	-	-	-	-	-
drift-turn	46 | 54 | 0 | 0	36 | 64 | 0 | 0	21 | 57 | 14 | 7	85 | 15 | 0 | 0	21 | 36 | 29 | 14	–	
	85 | 15 | 0 | 0	100 | 0 | 0 | 0	62 | 38 | 0 | 0	92 | 8 | 0 | 0	50 | 43 | 7 | 0	54 | 46 | 0 | 0	06:29
horsejump-low	64 | 36 | 0 | 0	79 | 21 | 0 | 0	69 | 31 | 0 | 0	100 | 0 | 0 | 0	50 | 50 | 0 | 0	–	
	77 | 23 | 0 | 0	85 | 15 | 0 | 0	54 | 46 | 0 | 0	100 | 0 | 0 | 0	62 | 38 | 0 | 0	7 | 64 | 29 | 0	06:37
parkour	68 | 29 | 4 | 0	100 | 0 | 0 | 0	90 | 7 | 3 | 0	86 | 14 | 0 | 0	28 | 59 | 10 | 3	–	
	-	-	-	-	-	-	-
scooter-gray	93 | 7 | 0 | 0	85 | 15 | 0 | 0	85 | 15 | 0 | 0	57 | 43 | 0 | 0	15 | 69 | 15 | 0	–	
	92 | 8 | 0 | 0	100 | 0 | 0 | 0	85 | 15 | 0 | 0	85 | 15 | 0 | 0	31 | 62 | 8 | 0	15 | 62 | 23 | 0	04:39
stroller	31 | 59 | 10 | 0	93 | 7 | 0 | 0	55 | 41 | 3 | 0	93 | 7 | 0 | 0	14 | 72 | 10 | 3	–	
	-	-	-	-	-	-	-
surf	79 | 21 | 0 | 0	86 | 10 | 3 | 0	93 | 7 | 0 | 0	100 | 0 | 0 | 0	0 | 34 | 62 | 3	–	
	-	-	-	-	-	-	-
swing	69 | 31 | 0 | 0	92 | 8 | 0 | 0	62 | 38 | 0 | 0	85 | 15 | 0 | 0	8 | 23 | 23 | 46	–	
	100 | 0 | 0 | 0	100 | 0 | 0 | 0	100 | 0 | 0 | 0	100 | 0 | 0 | 0	7 | 86 | 7 | 0	54 | 46 | 0 | 0	04:16
dweebs	92 | 8 | 0 | 0	92 | 8 | 0 | 0	100 | 0 | 0 | 0	100 | 0 | 0 | 0	46 | 54 | 0 | 0	–	
	100 | 0 | 0 | 0	100 | 0 | 0 | 0	100 | 0 | 0 | 0	92 | 8 | 0 | 0	85 | 15 | 0 | 0	38 | 23 | 38 | 0	05:53

Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. 2023. AMT: All-Pairs Multi-Field Transforms for Efficient Frame Interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9801–9810.

Chunxu Liu, Guozhen Zhang, Rui Zhao, and Limin Wang. 2024. Sparse Global Matching for Video Frame Interpolation with Large Motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 19125–19134. https://doi.org/10.1109/CVPR52733.2024.01809

Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. 2022. Video Frame Interpolation With Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3532–3542.

Simon Niklaus and Feng Liu. 2020. Softmax Splatting for Video Frame Interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5437–5446.

Junheum Park, Chul Lee, and Chang-Su Kim. 2021. Asymmetric Bilateral Motion Estimation for Video Frame Interpolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14539–14548.

F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. 2016. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *Computer Vision and Pattern Recognition*.

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. 2022. FILM: Frame Interpolation for Large Motion. In *European Conference on Computer Vision (ECCV)*.

Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. 2021. XVFI: Extreme Video Frame Interpolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14489–14498.

Guangyang Wu, Xin Tao, Changlin Li, Wenyi Wang, Xiaohong Liu, and Qingqing Zheng. 2024. Perception-Oriented Video Frame Interpolation via Asymmetric Blending. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2753–2762.

Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. 2019. Video Enhancement with Task-Oriented Flow. *International Journal of Computer Vision (IJCV)* 127, 8 (2019), 1106–1125.

Guozhen Zhang, Chunxu Liu, Yutao Cui, Xiaotong Zhao, Kai Ma, and Limin Wang. 2024. VFIMamba: Video Frame Interpolation with State Space Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. 2023. Extracting Motion and Appearance via Inter-Frame Attention for Efficient Video Frame Interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5682–5692.

Zhihang Zhong, Gurunandan Krishnan, Xiao Sun, Yu Qiao, Sizhuo Ma, and Jian Wang. 2025. Clearer Frames, Anytime: Resolving Velocity Ambiguity in Video Frame Interpolation. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Vol. 15091. Springer Nature Switzerland, Cham, 346–363. https://doi.org/10.1007/978-3-031-73414-4_20