LC-Mamba: Local and Continuous Mamba with Shifted Windows for Frame Interpolation

Supplementary Material

A. Appendix

A.1. Additional Details on Ablation Studies

Tables 6 and 7 present the PSNR/SSIM and LPIPS/FloLPIPS metrics, respectively, illustrating the impact of window size and the use of shifted windows on the performance of Hilbert curve-based scanning in our Ours-B model. The settings vary in window sizes of 4, 8, and 16, with and without shifted windows.

We observe that a window size of 8 provides balanced performance across both low and high resolutions. Specifically, on low-resolution datasets like Vimeo90K, omitting shifted windows slightly improved performance. For instance, using a window size of 8 without shifted windows yields a PSNR/SSIM of 36.45/0.9813, compared to 36.43/0.9813 with shifted windows. Conversely, at 4K resolution (Xiph-4K dataset), including shifted windows enhances global information extraction, resulting in better performance (PSNR/SSIM of 34.26/0.9046 with shifted windows versus 34.15/0.9042 without).

When the window size is increased to 16, the model effectively captures global features even without shifted windows, maintaining or slightly improving performance levels. For example, with a window size of 16 without shifted windows, the PSNR/SSIM on the Xiph-4K dataset is 34.23/0.9045, comparable to smaller window sizes with shifted windows.

Table 6. Ablation studies for window settings. The "Settings" column shows window size and whether shifting is used, while the other columns display performance (PSNR/SSIM).

Settings	Vimeo90K	Xiph-2K	Xiph-4K	SNU-FILM (avg.)
4 w/ shift	36.45 / **0.9813**	**36.98 / 0.9455**	**34.26 / 0.9052**	**33.08 / 0.9431**
4 w/o shift	36.44 / **0.9813**	36.90 / 0.9452	34.23 / 0.9046	33.05 / 0.9430
8 w/ shift	36.43 / **0.9813**	36.90 / 0.9452	**34.26** / 0.9046	33.02 / 0.9429
8 w/o shift	36.45 / **0.9813**	36.78 / 0.9448	34.15 / 0.9042	32.96 / 0.9428
16 w/ shift	36.44 / **0.9813**	36.88 / 0.9454	34.15 / 0.9047	33.01 / 0.9429
16 w/o shift	**36.46 / 0.9813**	36.88 / 0.9449	34.23 / 0.9045	33.05 / 0.9429

Table 7 shows the perceptual metrics (LPIPS/FloLPIPS), which further confirm these observations. Larger window sizes cover wider areas, enhancing the integration of spatiotemporal information; however, excessively large windows may have limitations in capturing fine motion details, as indicated by slightly higher LPIPS values.

Overall, a window size of 8 strikes a good balance between capturing global and local information. The inclusion of shifted windows is more beneficial at higher resolutions, enhancing global information extraction. These findings demonstrate the importance of selecting appropriate window sizes and configurations to optimize performance across different resolutions.

Table 7. Ablation studies for window settings. The "Settings" column shows window size and whether shifting is used, while the other columns display performance (LPIPS/FloLPIPS).

Settings	Vimeo90K	Xiph-2K	Xiph-4K	SNU-FILM (avg.)
4 w/ shift	0.0208 / **0.0380**	0.1060 / 0.1388	0.2323 / 0.2631	0.0607 / 0.0984
4 w/o shift	**0.0207** / 0.0381	0.1058 / 0.1386	**0.2314** / 0.2623	**0.0605 / 0.0982**
8 w/ shift	0.0208 / **0.0380**	**0.1056 / 0.1378**	0.2316 / **0.2609**	0.0612 / 0.0997
8 w/o shift	0.0210 / 0.0384	0.1059 / 0.1381	0.2332 / 0.2638	0.0611 / 0.0993
16 w/ shift	0.0211/ 0.0388	0.1061 / 0.1391	0.2339 / 0.2635	0.0621 / 0.1005
16 w/o shift	0.0208 /0.0381	0.1057 / **0.1378**	0.2322 / 0.2632	**0.0605** / 0.0988

B. Evaluating Temporal Consistency in Local and Global Frame Interpolation

We compare our LC-Mamba model against recent state-of-the-art video frame interpolation methods, including CNN-based approaches [1, 6, 11, 15, 32], Transformer-based models [19, 26, 46], and Mamba-based approaches [47], across multiple benchmark datasets.

As shown in Table 8, our method effectively handles both local and global motions by varying the frame interval (1–3) on the Vimeo90K dataset as part of a multi-resolution training strategy. Consequently, it achieves superior performance on both low- and high-resolution datasets without increasing model complexity, highlighting its efficiency and effectiveness. Furthermore, we categorize our model into three sizes—Compact (C), Efficient (E), and Balance (B)—by varying N_1, N_2, and channels. C and E use 16 channels ($N_1 = N_2 = 2$ for C; $N_1 = N_2 = 4$ for E), while B uses 32 channels with $N_1 = N_2 = 2$. All variants use a fixed window size of 8 with window shifting.

C. Additional Visual Results

Figures 6, 7, and 8 illustrate that our window-based hierarchical architecture and Hilbert curve-based scanning effectively maintain temporal consistency, generating visually compelling results for videos exhibiting both local and global motions. By simultaneously modeling detailed local motion and overall global scene dynamics, our method synthesizes stable, realistic video sequences.

Additional qualitative comparisons with state-of-the-art methods (VFIFormer [26], EMA [46], SGM-VFI [19], VFI-Mamba [47]) are shown in Figures 9 and 10. These results demonstrate that our method effectively synthesizes both fine, small-scale and large motions.

Table 8. Additional quantitative comparison across benchmarks (IE for Middlebury; PSNR/SSIM for Vimeo90K, UCF101, Xiph, and SNU-FILM). The best and second-best results are highlighted in **bold** and underlined, respectively. "Out of Memory" is denoted as "OOM," and "†" indicates our own test results; other results are cited from [11, 14, 15, 26, 35, 46]. Evaluation procedures followed those of [14] for Vimeo90K, UCF101, and Middlebury, [30] for Xiph, and [15] for SNU-FILM, with Test-Time Augmentation (TTA) disabled.

Method	Vimeo90K	UCF101	Xiph		M.B.	SNU-FILM				Params (M)
			2K	4K		Easy	Medium	Hard	Extreme	
ToFlow [1]	33.73/0.9682	34.58/0.9667	33.93/0.922	30.74/0.856	2.15	39.08/0.9890	34.39/0.9740	28.44/0.9180	23.39/0.8310	1.4
IFRNet [15]	35.80/0.9794	35.29/0.9693	36.00/0.936	33.99/0.893	1.95	40.03/0.9905	35.94/0.9793	30.41/0.9358	25.05/0.8587	5
M2M [11]	35.47/0.9778	35.28/0.9694	36.44/0.943	33.92/0.899	2.09	39.66/0.9904	35.74/0.9794	30.30/0.9360	25.08/0.8604	7.6
SoftSplat [30]	36.10/0.9802	35.39/0.9697	36.62/0.944	33.60/0.901	**1.81**	39.88/0.9897	35.68/0.9772	30.19/0.9312	24.83/0.8500	7.7
RIFE [14]	35.61/0.9779	35.28/0.969	36.19/0.938	33.76/0.894	1.96	39.80/0.9903	35.76/0.9787	30.36/0.9351	25.27/0.8601	9.8
BMBC [31]	35.01/0.9764	35.15/0.9689	32.82/0.928	31.19/0.880	2.04	39.90/0.9902	35.31/0.9774	29.33/0.9270	23.92/0.8432	11.1
EMA-S [46]	36.07/0.9794†	35.34/0.9696†	36.54/0.942†	34.24/0.902†	1.94†	39.81/0.9903†	35.88/0.9792†	30.68/0.9371†	25.47/0.8627†	14.5
VFIMamba-S [47]	36.09/0.9800†	35.35/0.9696†	36.71/0.942†	34.26/0.902†	1.97†	40.21/0.9912†	36.17/0.9802†	30.80/0.9382†	25.59/0.8655†	16.8
VFIFormer-S [26]	36.37/0.9810†	35.36/0.9698†	36.55/0.943†	33.37/0.899†	1.89†	40.02/0.9906†	35.91/0.9793†	30.22/0.9348†	24.80/0.8568†	17.1
ABME [32]	36.18/0.9805	35.38/0.9698	36.53/0.944	33.73/0.901	2.01	39.59/0.9901	35.77/0.9789	30.58/0.9364	25.42/0.8639	18.1
SGM-VFI-S-1/2 [19]	35.81/0.9785†	35.33/0.9692†	36.06/0.940†	33.26/0.897†	1.87†	40.36/0.9900†	36.12/0.9787†	30.62/0.9351†	25.38/0.8615†	20.8
SepConv [5]	33.79/0.9702	34.78/0.9669	34.77/0.929	32.06/0.880	2.27	39.41/0.9900	34.97/0.9762	29.36/0.9253	24.31/0.8448	21.7
AdaCoF [16]	34.47/0.9730	34.90/0.9680	34.86/0.928	31.68/0.870	2.24	39.80/0.9900	35.05/0.9754	29.46/0.9244	24.31/0.8439	21.8
DAIN [2]	34.71/0.9756	34.99/0.9683	35.95/0.940	33.49/0.895	2.04	39.73/0.9902	35.46/0.9780	30.17/0.9335	25.09/0.8584	24.0
VFIFormer [26]	36.50/**0.9815**†	35.42/0.9699†	OOM†	OOM†	1.82†	40.12/0.9907†	36.09/0.9798†	30.67/0.9378†	25.43/0.8643†	24.1
CAIN [6]	34.65/0.9730	34.91/0.9690	35.21/0.937	32.56/0.901	2.28	39.89/0.9900	35.61/0.9776	29.90/0.9292	24.78/0.8507	42.8
EMA [46]	36.50/0.9814†	35.38/0.9697†	36.74/0.944†	34.54/0.905†	1.84†	39.57/0.9905†	35.85/0.9797†	30.80/0.9389†	25.59/0.8650†	65.6
VFIMamba [47]	36.45/0.9807†	35.37/0.9699†	37.02/0.944†	34.39/0.904†	1.89†	**40.41**/0.9903†	**36.30**/0.9794†	30.89/0.9387†	25.68/0.8661†	66.1
Ours-C	36.10/0.9801	35.38/0.9700	37.12/0.946	34.81/0.908	1.94	40.10/**0.9915**	36.11/0.9809	30.81/0.9405	25.69/0.8710	4.3
Ours-E	36.20/0.9802	35.42/0.9699	37.17/0.946	34.99/0.910	1.96	40.15/0.9912	36.18/0.9809	30.89/0.9416	25.81/**0.8725**	6.7
Ours-B	**36.52**/0.9810	**35.47/0.9703**	**37.33/0.947**	**35.14/0.911**	1.90	40.20/0.9909	**36.30/0.9810**	**31.00/0.9417**	**25.83**/0.8722	16.2

Figure 6. Visualization of Overlay, Ground-Truth, Synthesis, Difference, Bidirectional Flow, and Mask on the Vimeo90K [43] dataset. "Overlay" denotes the overlay of the two input frames, "Ground-Truth" is the correct intermediate frame, "Synthesis" is the interpolated frame, and "Difference" represents the absolute error between the ground truth and synthesis. "Forward Flow and Backward Flow" depict the motion flows at time t for frames 0 and 1, respectively, while "Mask" is used to blend the frames warped by each flow.

Figure 7. Visualization of Overlay, Ground-Truth, Synthesis, Difference, Bidirectional Flow, and Mask on the Xiph [29] dataset. "Overlay" denotes the overlay of the two input frames, "Ground-Truth" is the correct intermediate frame, "Synthesis" is the interpolated frame, and "Difference" represents the absolute error between the ground truth and synthesis. "Forward Flow and Backward Flow" depict the motion flows at time t for frames 0 and 1, respectively, while "Mask" is used to blend the frames warped by each flow.

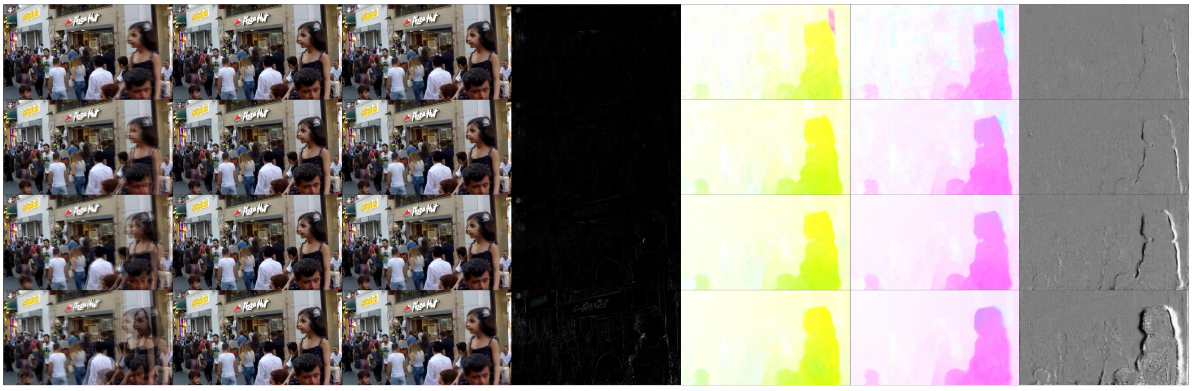

Figure 8. Visualization of Overlay, Ground-Truth, Synthesis, Difference, Bidirectional Flow, and Mask on the SNU-FILM [6] dataset. "Overlay" denotes the overlay of the two input frames, "Ground-Truth" is the correct intermediate frame, "Synthesis" is the interpolated frame, and "Difference" represents the absolute error between the ground truth and synthesis. "Forward Flow and Backward Flow" depict the motion flows at time t for frames 0 and 1, respectively, while "Mask" is used to blend the frames warped by each flow. Each row corresponds to a different difficulty level (Easy, Medium, Hard, Extreme).

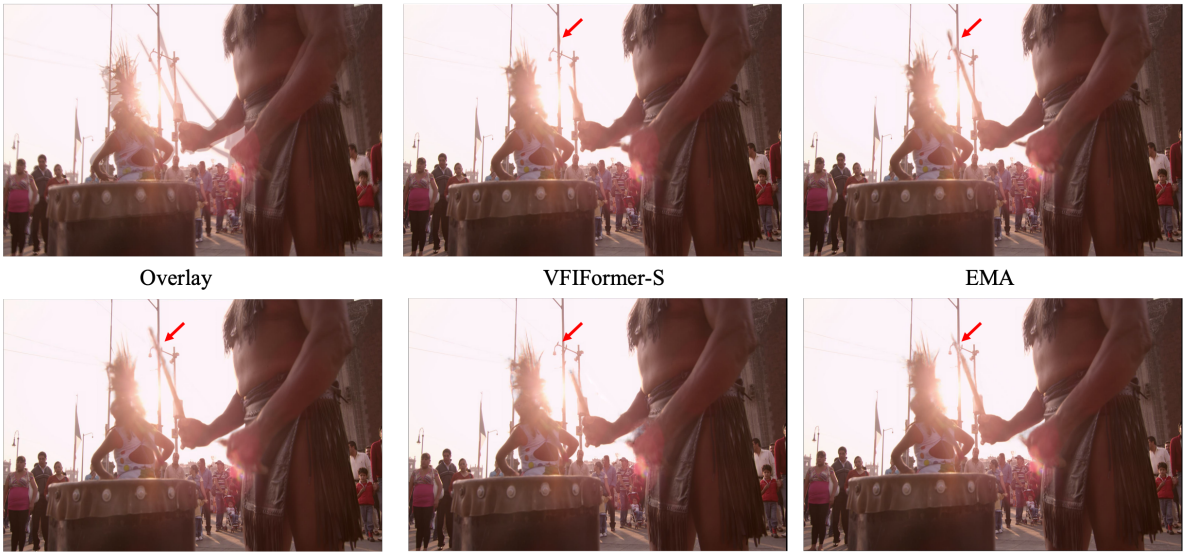

Figure 9. Visual comparison on the Xiph [29] dataset. Our Balance model better captures the fine details of the wooden stick striking the drum, as indicated by the red arrow. Note that "Overlay" denotes the overlay of the two input frames.

Figure 10. Visual comparison on Extreme levels of the SNU-FILM [6] dataset. The red arrows highlight regions with large motions and fine details (e.g., hair and finger features), resulting in smoother synthesis. Note that "Overlay" denotes the overlay of the two input frames.