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

Min Wu Jeong, Chae Eun Rhee
Hanyang University

Github

Introduction

Summary of Our Work

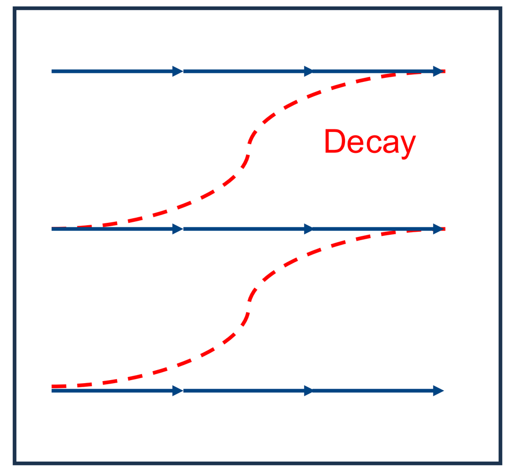
2D Hilbert Curve-based Selective State Scan
- **Maintains Spatial Correlation:** Preserves spatial relationships within and between windows.
- **Preserves Spatial Continuity:** Prevents distortion of spatial information during scanning.

Hierarchical Shifted Window
- **8x8 Restricted Scan Area:** Enables fine-grained observation of local spatial characteristics in high-resolution frames.
- **Reduced Information Decay:** Minimizes loss of historical information.
- **Multi-Scale Motion Capture:** Effectively captures complex motion between frames.

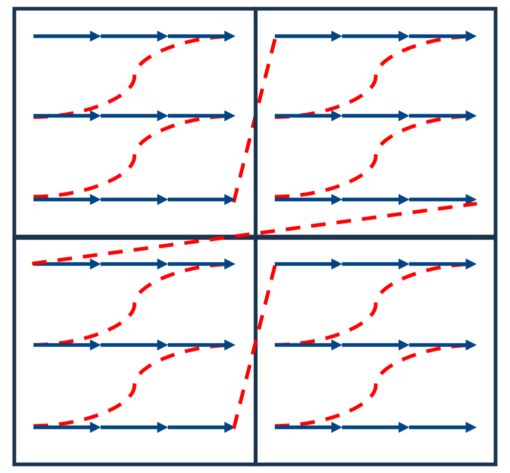
Temporal Scan
- **Leverages Interleaved Selective State Scan:** 2D Hilbert curve-based.
- **Simultaneous Spatiotemporal Modeling:** Enables integrated analysis of temporal and spatial relationships between two frames

Limitations of Previous Methods

- **CNN:** Difficulty in accurately capturing complex and large movements due to a **small receptive field**.
- **ViT:** Advantageous for modeling long-range dependencies, but incurs **high computational costs** for high-resolution and real-time processing due to the **quadratic complexity of self-attention**.
- **Mamba:** Suitable for high-resolution processing with linear computational complexity, but suffers from **loss of spatial characteristics** and **weakened information transfer between tokens (information decay)** during the 1D sequential scanning of 2D images, limiting local pixel relationship modeling.



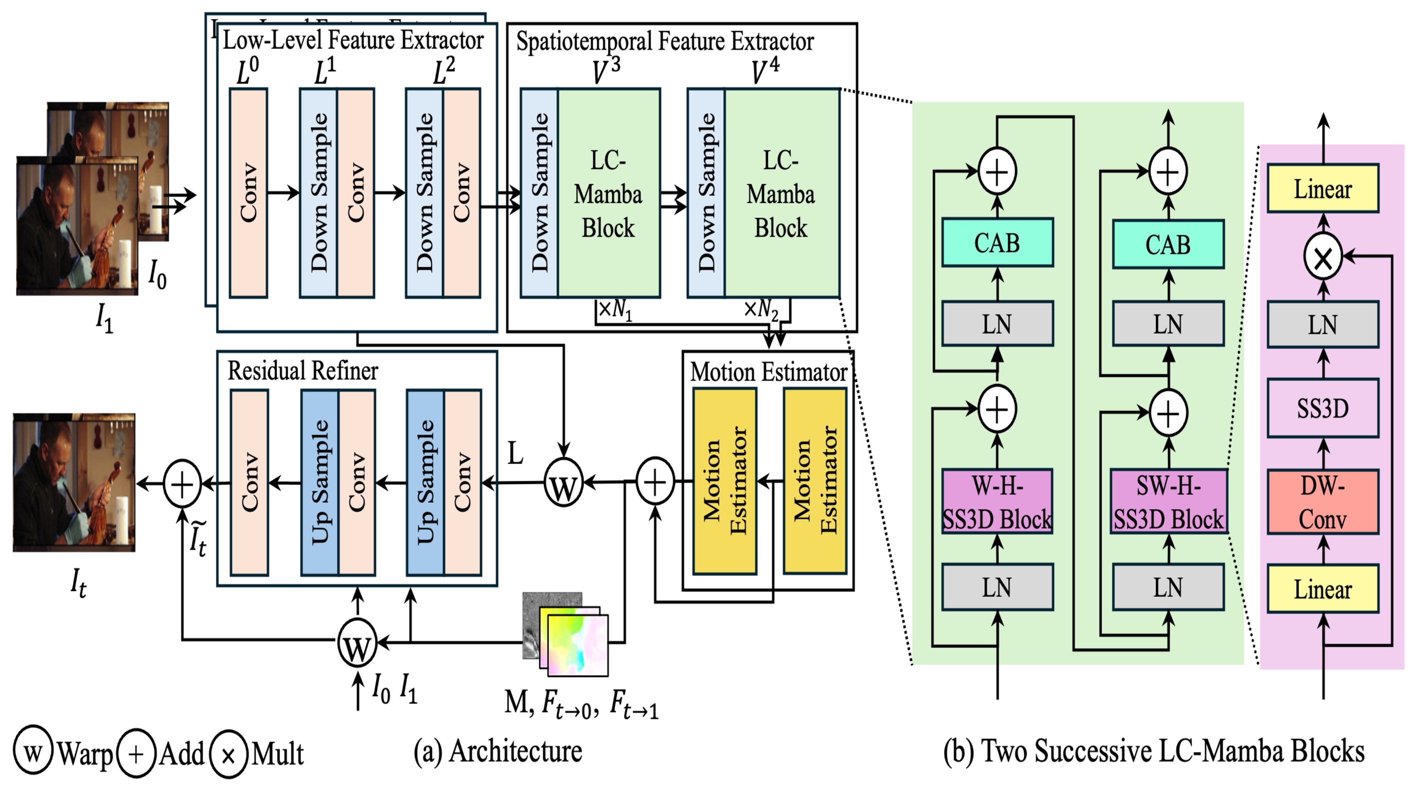
Sequential 1D Scan



Sequential Local 1D Scan

Proposed Method

Overall Architecture of LC-Mamba

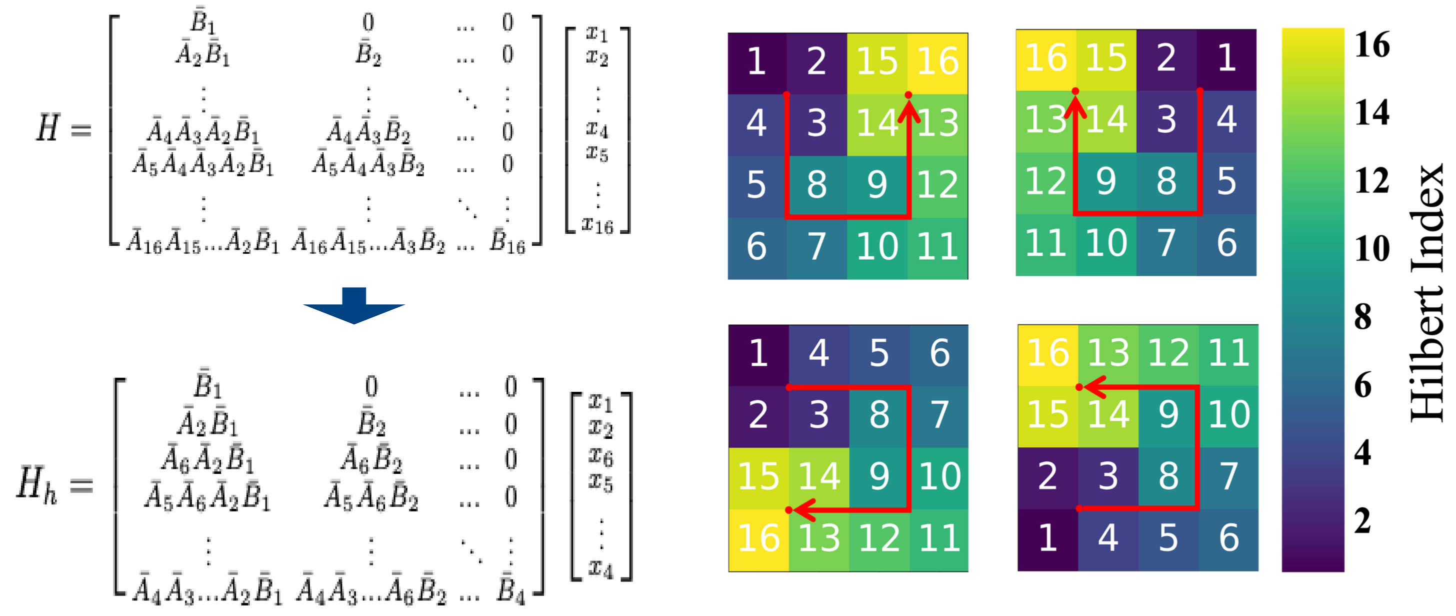


(a) Architecture

(b) Two Successive LC-Mamba Blocks

2D Hilbert Curve-based Selective State Scan

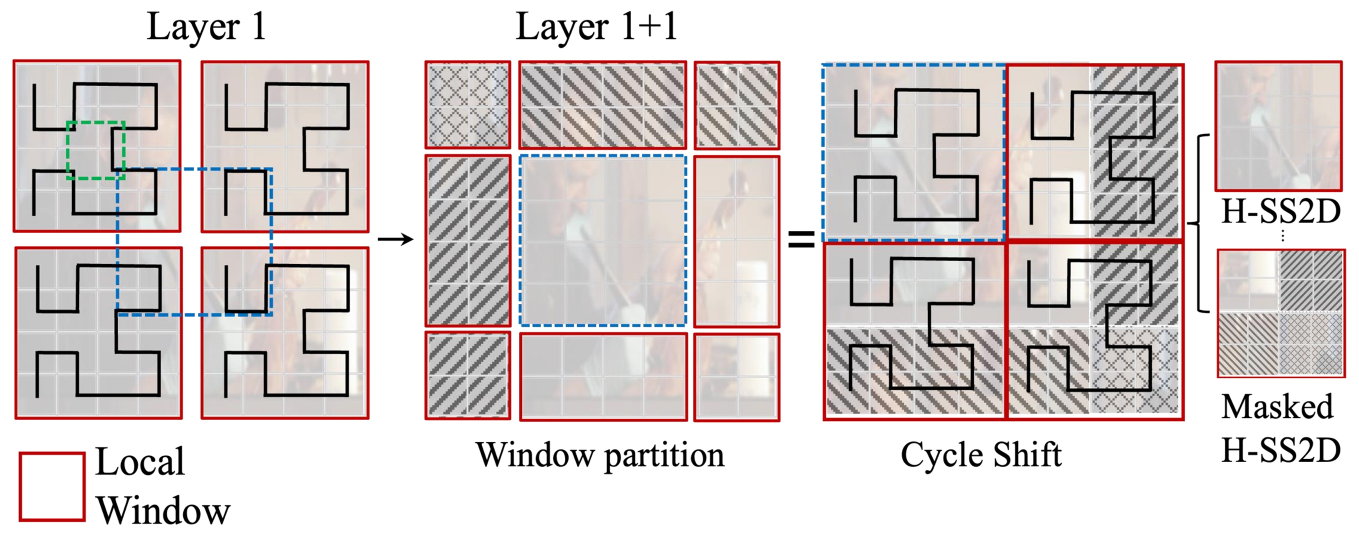
Preserves **pixel-level spatial adjacency** using **Hilbert curves** for 2D-to-1D conversion. Its **four-path Hilbert structure** enhances **information transfer efficiency** and **compensates for loss/decay** via alternative paths.



Hierarchical Shifted Window

Simultaneously secures **locality and global connectivity** through **shifted windows and Δ (Delta) gating**, precisely controlling information flow to effectively model spatial information.

$$h(t) = \bar{A}h(t-1) + \bar{B}x(t)$$
$$\bar{A} = e^{\Delta A}$$
$$\bar{B} = (\Delta A)^{-1}(e^{\Delta A} - I)\Delta B \approx \Delta B.$$
$$h(t) = \bar{A}h(t-1) \quad \bar{A} = 1, \bar{B} = 0 \quad \text{if } \Delta = 0$$



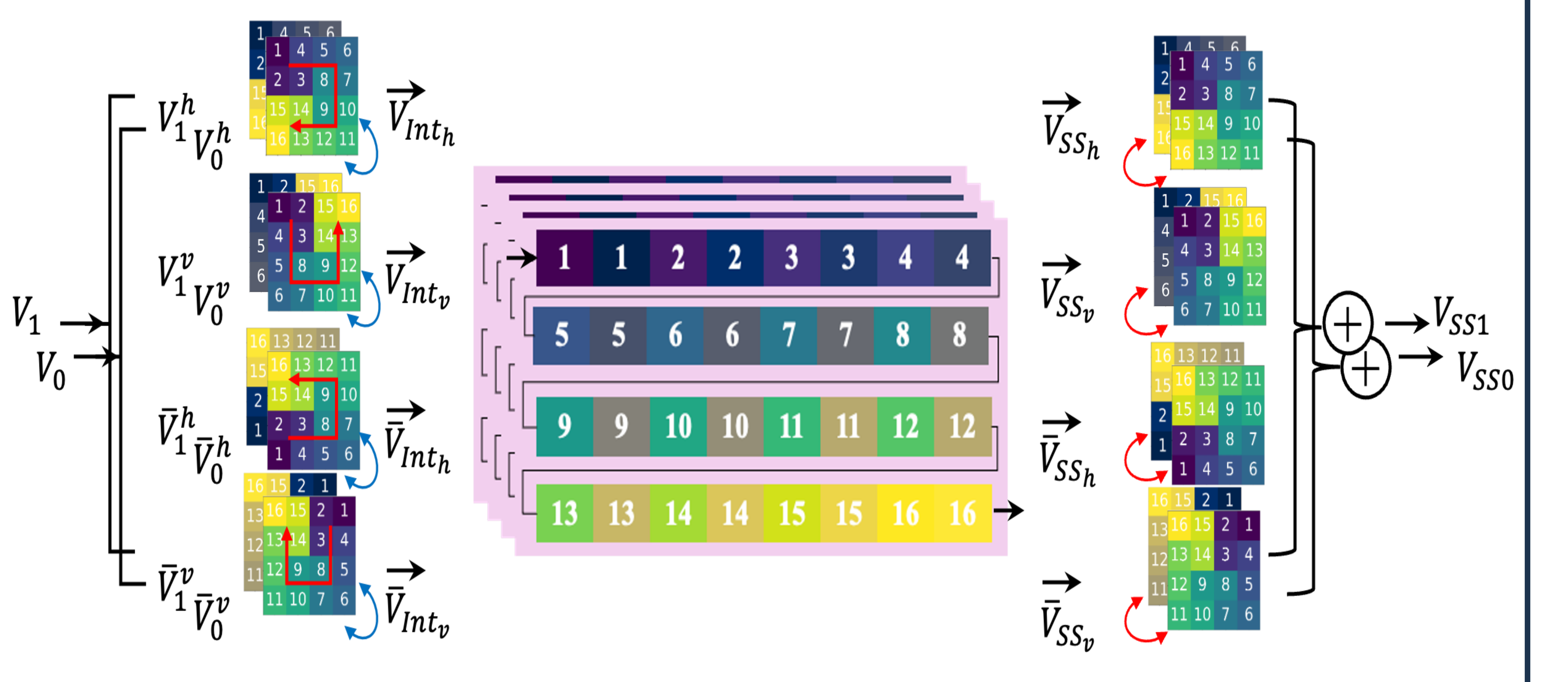
Local Window

Window partition

Cycle Shift

Temporal Scan

Effectively captures complex **spatiotemporal correlations** between frames through sophisticated **spatiotemporal relationship modeling** based on **interleaved H-SS2D scanning**.



Rearrange and Interleaving

Selective Scan

Revert and Deinterleaving

Experiments

Quantitative Result

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 2K	Xiph 4K	M.B.	SNU-FILM Easy	SNU-FILM Medium	SNU-FILM Hard	SNU-FILM Extreme	Params (M)	Flops (T)
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	0.62
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	0.21
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	0.26
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	0.94
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	0.20
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	2.50
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	0.39
VFIMamba-S [47]	36.09/0.9800†	35.35/0.9696†	36.71/0.942†	34.26/0.902†	1.97†	40.21/<u>0.9912</u>†	36.17/0.9802†	30.80/0.9382†	25.59/0.8655†	16.8	0.39
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	1.30
SGM-VFI-S-1/2 [19]	35.81/0.9785†	35.33/0.9692†	36.06/0.940†	33.26/0.897†	1.87†	<u>40.36</u>/0.9900†	36.12/0.9787†	30.62/0.9351†	25.38/0.8615†	20.8	1.96
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	0.38
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	0.36
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	5.51
VFIFormer [26]	<u>36.50</u>/**0.9815**†	<u>35.42</u>/0.9699†	OOM†	OOM†	<u>1.82</u>†	40.12/0.9907†	36.09/0.9798†	30.67/0.9378†	25.43/0.8643†	24.1	47.41
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	1.29
EMA [46]	36.50/<u>0.9814</u>†	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	1.64
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†	<u>30.89</u>/0.9387†	25.68/0.8661†	66.1	1.54
Ours-C	36.10/0.9801	35.38/<u>0.9700</u>	37.12/<u>0.946</u>	34.81/0.908	1.94	40.10/**0.9915**	36.11/<u>0.9809</u>	30.81/0.9405	25.69/0.8710	4.3	0.27
Ours-E	36.20/0.9802	<u>35.42</u>/0.9699	<u>37.17</u>/<u>0.946</u>	<u>34.99</u>/<u>0.910</u>	1.96	40.15/<u>0.9912</u>	<u>36.18</u>/<u>0.9809</u>	<u>30.89</u>/<u>0.9416</u>	<u>25.81</u>/**0.8725**	6.7	0.29
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**/<u>0.8722</u>	16.2	1.07

Ablation Study

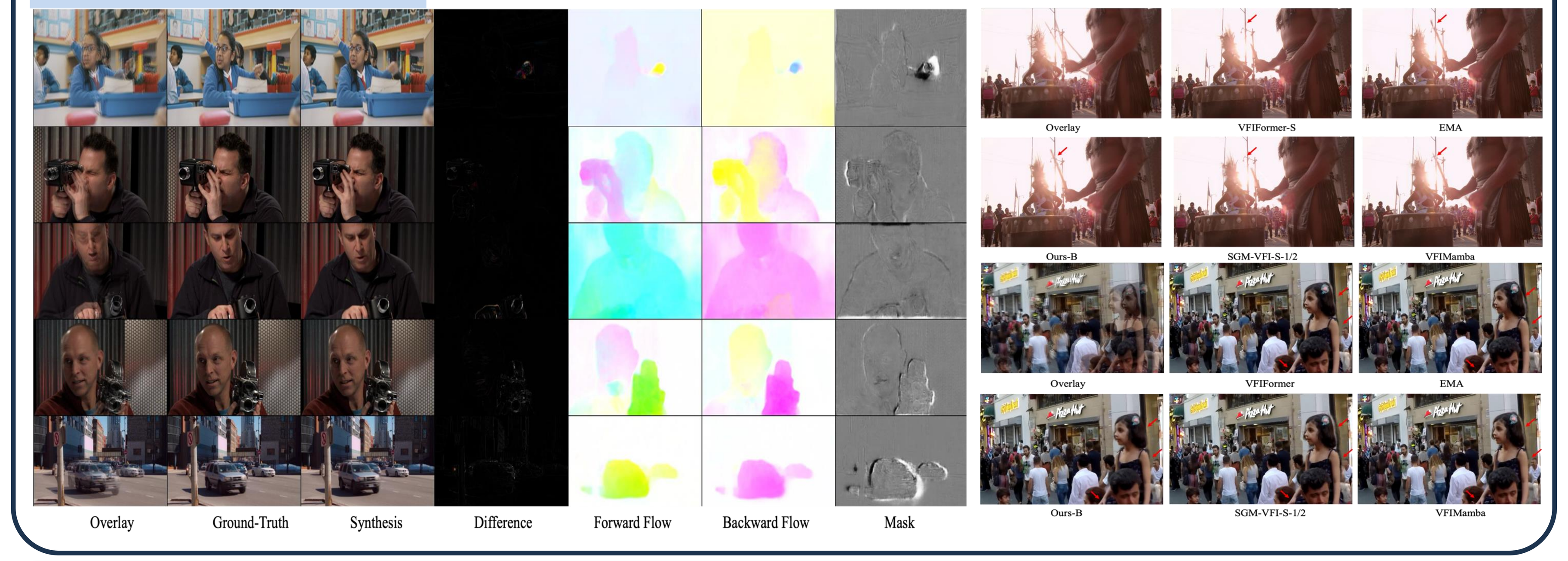
Table 3. Performance comparison of different scanning methods

Scanning	Vimeo90K	Xiph-2K	Xiph-4K	SNU-FILM(avg.)
Bidirection w/ ILV	35.41/0.9799	36.00/0.9381	33.13/0.8937	32.32/0.9405
Cross w/ ILV	36.07/0.9799	35.73/0.9362	33.80/0.8947	32.53/0.9413
Continuous w/ ILV	36.09/0.9800	<u>36.57</u>/<u>0.9428</u>	33.99/<u>0.9010</u>	24.59/0.8335
Local w/ ILV	36.11/<u>0.9801</u>	36.38/0.9415	<u>34.01</u>/0.9008	<u>32.62</u>/0.9411
Z-order w/ ILV	<u>36.13</u>/0.9800	35.91/0.9371	33.30/0.8932	32.36/<u>0.9417</u>
SW-H-SS3D	**36.19/0.9803**	**36.67/0.9437**	**34.26/0.9036**	**32.89/0.9426**

Table 4. 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.)
8 w/ shift	36.43/**0.9813**	**36.90**/<u>0.9452</u>	**34.26**/<u>0.9046</u>	<u>33.02</u>/**0.9429**
8 w/o shift	<u>36.45</u>/**0.9813**	36.78/0.9448	34.15/0.9042	32.95/<u>0.9428</u>
16 w/ shift	36.44/**0.9813**	<u>36.88</u>/**0.9454**	34.15/**0.9047**	<u>33.02</u>/**0.9429**
16 w/o shift	**36.46/0.9813**	<u>36.88</u>/0.9449	<u>34.23</u>/0.9045	**33.05/0.9429**

Qualitative Result



Conclusion

- **Enhanced Mamba for VFI:** Hilbert curves & shift-window scanning improve spatial continuity and balance local/global information.
- **Achieves higher efficiency** in low/high-resolution settings and effectively captures diverse motion patterns.