MoTIF: Learning Motion Trajectories with Local Implicit Neural Functions for Continuous Space-Time Video Super-Resolution
Supplementary Materials

Yi-Hsin Chen* Si-Cun Chen* Yi-Hsin Chen Yen-Yu Lin Wen-Hsiao Peng
National Yang Ming Chiao Tung University, Taiwan
{yhchen12101, sicun.mapl, karta6120}.cs09@nycu.edu.tw
lin@cs.nycu.edu.tw wpeng@cs.nctu.edu.tw

This document provides additional results for

- More comparison with the F-STVSR methods in Section A1;
- Replacing Raft-lite in MoTIF with other pre-trained flow estimation network in Section A2;
- Fourier analysis of forward and backward motion in Section A3;
- Subjective comparison in Section A4;
- Implementation details in Section A5.

A1. More Comparisons with F-STVSR Methods

This experiment compares our MoTIF with the state-of-the-art F-STVSR methods. Similar comparison is provided in the main paper, following the setting of VideoINR [2], in which the training is done on Adobe240fps [10] dataset and with 2 reference frames. Here, we follow the common test protocol [14, 15] of the F-STVSR task to perform training with 4 reference frames.

In the present case, we have access to $I^L_{-1}, I^L_0, I^L_1, I^L_2$, in generating a high-resolution video frame $I^H_t, t \in [-1, 2]$. To extend our scheme to 4 reference frames, (1) we follow ZSM [14] to generate the reference features $F^L_{-1}, F^L_0, F^L_1, F^L_2$ and the intermediate features $F^L_{(-1,0)}, F^L_{(0,1)}, F^L_{(1,2)}$. (2) We then have the motion latent T^L_0 encode jointly information from multiple pairs $\{M^L_{0\to i}, Z^L_{0\to i}\}, i = -1, 1, 2$ of the forward flow map $M_{0\to i}$ and its reliability map $Z^L_{0\to i}$, with i referring to the reference frames except I^L_0. The same process is repeated to generate the other motion latents $T^L_i, i = -1, 1, 2$. (3) Based on these motion latents, we aggregate features F^H_i from the 4 reference frames to synthesize F^H_t, Z^H_t. (4) During the decoding of the RGB values, the intermediate feature is chosen from $F^L_{(-1,0)}, F^L_{(0,1)}, F^L_{(1,2)}$ depending on which interval t sits in. For example, if $t = -0.3$, the intermediate feature is $F^L_{(-1,0)}$, and if $t = 1.8$, the intermediate feature is $F^L_{(1,2)}$.

From Table A1, we see that our MoTIF, although trained for the C-STVSR task, shows comparable performance to RSTT-L [3] and TMNet [15] on the F-STVSR task. Both RSTT-L [3] and TMNet [15] are the state-of-the-art one-stage F-STVSR methods. They, however, are not able to support the C-STVSR task. VideoINR [2] is not included in this comparison since it accepts only 2 reference frames.

A2. Raft-lite vs. PWC-Net in MoTIF

Following the same experimental setup in Section A1, Table A2 provides additional results by replacing Raft-lite [12] in MoTIF with the pre-trained PWC-Net [11]. As shown, the change in PSNR/SSIM is minor. This indicates that our MoTIF can work well with well-behaved, off-the-shelf flow estimation networks.

A3. Fourier Analysis Results

Figs. A1 and A2 analyze the signal spectra of the forward and backward motion representations. We take a vertical slice of pixels in the first columns of Fig. A1 and A2 as examples, and represent their forward or backward motion over 33 consecutive video frames as functions of time. At each vertical pixel location, we conduct 1-D Fourier transform of the motion signal along the temporal dimension. In each figure, (1) the first column superimposes the first and the last frames of the test sequence, (2) the second column shows the forward motion from the first frame to the last frame, and (3) the third and the fourth columns visualize the spectra of the forward and backward motion, respectively.

In Fig. A1, at each spatial location, the 1-D Fourier transform along the temporal dimension is applied to the hor-

*Both authors contributed equally to this work.

Table A1: Performance comparison on the F-STVSR task. Red, blue, and **bold** indicate the best, the second best, and the third best performance, respectively. Quality metrics: PSNR/SSIM. Our MoTIF, although trained for the C-STVSR task, shows comparable performance to RSTT-L and TMNet on the F-STVSR task. Both RSTT-L and TMNet are the state-of-the-art one-stage F-STVSR methods. They are not able to support the C-STVSR task. See Section A1.

VFI Method	VSR Method	Vid4 [7]	Vimeo-Fast [16]	Vimeo-Medium [16]	Vimeo-Slow [16]
SuperSloMo[6]	Bicubic	22.84 / 0.5772	31.88 / 0.8793	29.94 / 0.8477	28.37 / 0.8102
SuperSloMo[6]	RCAN[17]	23.80 / 0.6397	34.52 / 0.9076	32.50 / 0.8884	30.69 / 0.8624
SuperSloMo[6]	RBPN[4]	23.76 / 0.6362	34.73 / 0.9108	32.79 / 0.8930	30.48 / 0.8584
SuperSloMo[6]	EDVR[13]	24.40 / 0.6706	35.05 / 0.9136	33.85 / 0.8967	30.99 / 0.8673
SepConv[9]	Bicubic	23.51 / 0.6273	32.27 / 0.8890	30.61 / 0.8633	29.04 / 0.8290
SepConv[9]	RCAN[17]	24.92 / 0.7236	34.97 / 0.9195	33.59 / 0.9125	32.13 / 0.8967
SepConv[9]	RBPN[4]	26.08 / 0.7751	35.07 / 0.9238	34.09 / 0.9229	32.77 / 0.9090
SepConv[9]	EDVR[13]	25.93 / 0.7792	35.23 / 0.9252	34.22 / 0.9240	32.96 / 0.9112
DAIN[1]	Bicubic	23.55 / 0.6268	32.41 / 0.8910	30.67 / 0.8636	29.06 / 0.8289
DAIN[1]	RCAN[17]	25.03 / 0.7261	35.27 / 0.9242	33.82 / 0.9146	32.26 / 0.8974
DAIN[1]	RBPN[4]	25.96 / 0.7784	35.55 / 0.9300	34.45 / 0.9262	32.92 / 0.9097
DAIN[1]	EDVR[13]	26.12 / 0.7836	35.81 / 0.9323	34.66 / 0.9281	33.11 / 0.9119
STARnet[5]		26.06 / 0.8046	36.19 / 0.9368	34.86 / 0.9356	33.10 / 0.9164
Zooming SlowMo[14]		26.31 / 0.7976	**36.81 / 0.9415**	35.41 / 0.9361	33.36 / 0.9138
TMNet[15]		26.43 / 0.8016	37.04 / 0.9435	35.60 / 0.9380	33.51 / 0.9159
RSTT-L[3]		26.43 / 0.7994	36.80 / 0.9403	35.66 / 0.9381	33.50 / **0.9147**
Ours		26.43 / **0.8013**	36.88 / 0.9427	**35.53 / 0.9372**	**33.46** / 0.9148

Table A2: PSNR/SSIM comparison of the pre-trained Raft [12] and PWC-Net [11] in MoTIF.

Flow Estimator	Vid4	Vimeo-Fast	Vimeo-Medium	Vimeo-Slow
Raft-lite [12]	26.43 / 0.8013	36.88 / 0.9427	35.53 / 0.9372	33.46 / 0.9148
PWC-Net [11]	26.40 / 0.8001	36.89 / 0.9432	35.52 / 0.9366	33.48 / 0.9161

izontal component (namely, the x-component) of the displacement vectors. The spectra shown are magnitude responses. We see that forward motion usually has much stronger responses in the low-frequency bands, especially the DC band (temporal frequency=0), than backward motion. On the other hand, backward motion has more high-frequency responses. This implies that the back motion representation is typically less smooth temporally.

In Fig. A2, a similar analysis is conducted on the vertical component (namely, the y-component) of the displacement vectors. Interestingly, both the forward and backward motion representations have similar frequency responses. This may be because most video sequences have less and smaller vertical motion.

A4. More Qualitative Results

Figs. A3 , A4 , A5 , and A6 provide more subjective quality comparisons. Our MoTIF preserves more high-frequency details than the other competing methods in tests with both in-distribution and out-of-distribution temporal scaling factors (cf. the buildings in Fig. A3, the heads of the ducks in Fig. A3, the edge of the butterfly in Fig. A4, the paper posted on the door of the train in Fig. A4, the license plate of the taxi in Fig. A5, and the legs of the race horse in Fig. A6).

A5. Implementation Details

A5.1. Reliability Maps

Following [8], we quantify the reliability of a forward optical flow map based on (1) the intensity warping error $Z^{int}_{0\to1}$, (2) the flow warping error $Z^{flow}_{0\to1}$, and (3) the local variances of the flow map. Consider $M^L_{0\to1}$ as an example. These metrics are given, respectively, by

$$Z^{int}_{0\to1} = \|I^L_0 - \omega(I^L_1, M^L_{0\to1})\|, \tag{1}$$

$$Z^{flow}_{0\to1} = \|M^L_{0\to1} - (-\omega(M^L_{1\to0}, M^L_{0\to1}))\|, \tag{2}$$

$$Z^{var}_{0\to1} = \sqrt{G((M^L_{0\to1})^2) - G(M^L_{0\to1})^2}, \tag{3}$$

where $\omega(A, B)$ denotes the operation of backward warping A based on B, e.g. $I^L_0 - \omega(I^L_1, M^L_{0\to1}) \equiv I^L_0(p) - I^L_1(p + M^L_{0\to1}(p)),\ \forall p$, with p denoting the pixel coordinates in I^L_0, and $G(\cdot)$ denotes the 3×3 Gaussian kernel. From Eq. (1), the intensity warping error evaluates the prediction error of I^L_0 by backward warping I^L_1 using $M^L_{0\to1}$. The flow warping error in Eq. (2) checks the consistency between $M^L_{0\to1}$ and $M^L_{1\to0}$. It is defined as the prediction error of $M^L_{0\to1}$ by backward warping $M^L_{1\to0}$ using $M^L_{0\to1}$. The sign flipping $-\omega(M^L_{1\to0}, M^L_{0\to1})$ accounts for the difference between $M^L_{0\to1}$ and $M^L_{1\to0}$ in their directions.

A5.2. Network Architecture

We further illustrate details of our network architecture in Fig. A7 and Fig. A8. As shown in Fig. A7, our motion encoder takes N group of motion features as input, where

N is the number of motion samples we use. Each motion feature includes the forward motion, the reliability map and two constant maps describing the source time and destination time of the forward motion, respectively.

References

[1] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] Zeyuan Chen, Yinbo Chen, Jingwen Liu, Xingqian Xu, Vidit Goel, Zhangyang Wang, Humphrey Shi, and Xiaolong Wang. Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2047–2057, 2022.

[3] Zhicheng Geng, Luming Liang, Tianyu Ding, and Ilya Zharkov. Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17441–17451, 2022.

[4] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Space-time-aware multi-resolution video enhancement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[6] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Ce Liu and Deqing Sun. A bayesian approach to adaptive video super resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 209–216, 2011.

[8] Simon Niklaus, Ping Hu, and Jiawen Chen. Splatting-based synthesis for video frame interpolation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[9] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[10] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)*, 2020.

[13] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[14] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P. Allebach, and Chenliang Xu. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[15] Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. Temporal modulation network for controllable space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6388–6397, 2021.

[16] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127(8):1106–1125, 2019.

[17] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

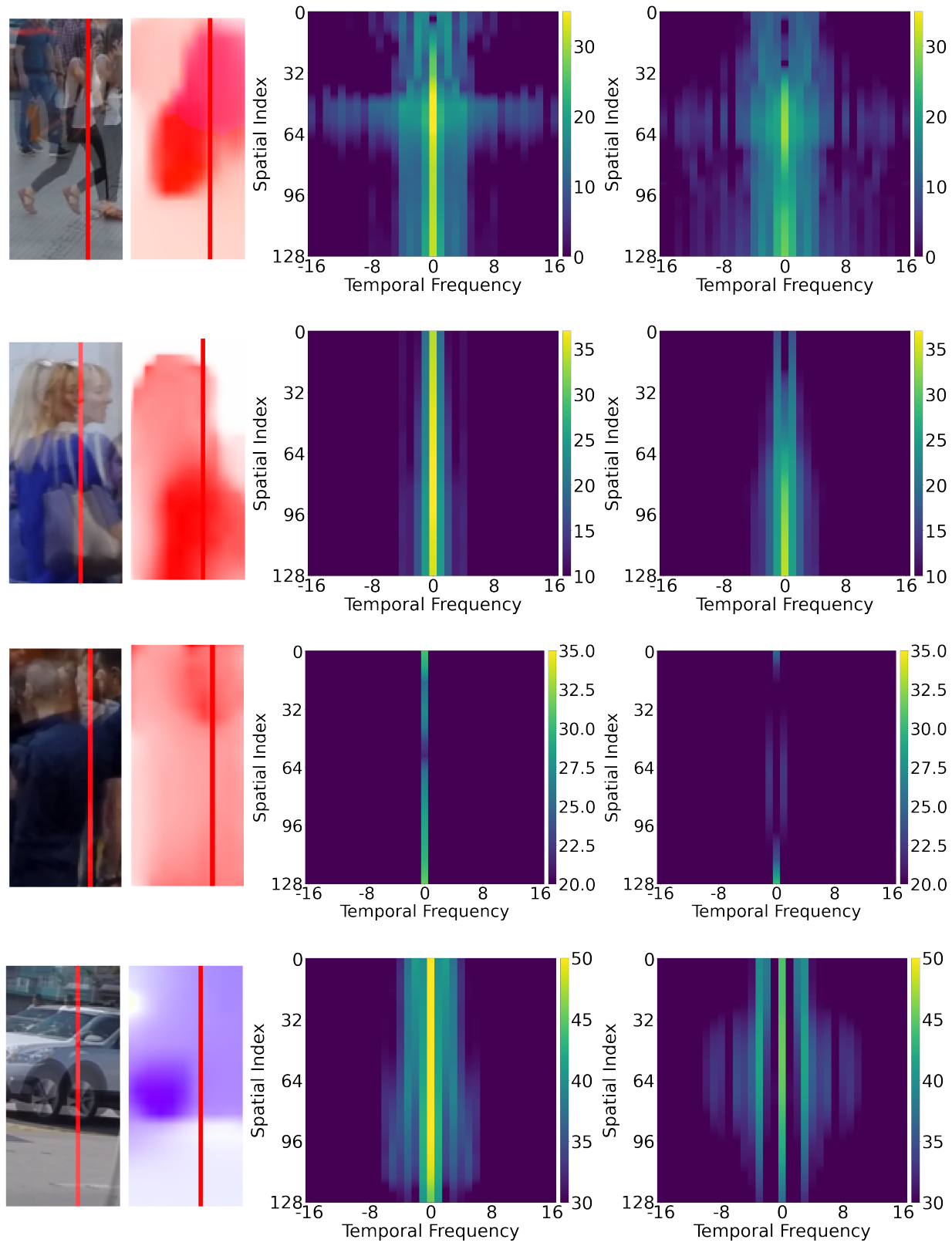

Figure A1: Fourier analysis of forward and backward motion. The first column shows the slice of pixels whose forward/backward motion are analyzed. The second column is the forward optical flow map. The third column is the temporal signal spectra of the horizontal components of the forward displacement vectors. The forth column is the temporal signal spectra of the horizontal component of the backward displacement vectors. The spectra shown are magnitude responses. Forward motion usually has much stronger responses in the low-frequency bands than backward motion. See Section A3.

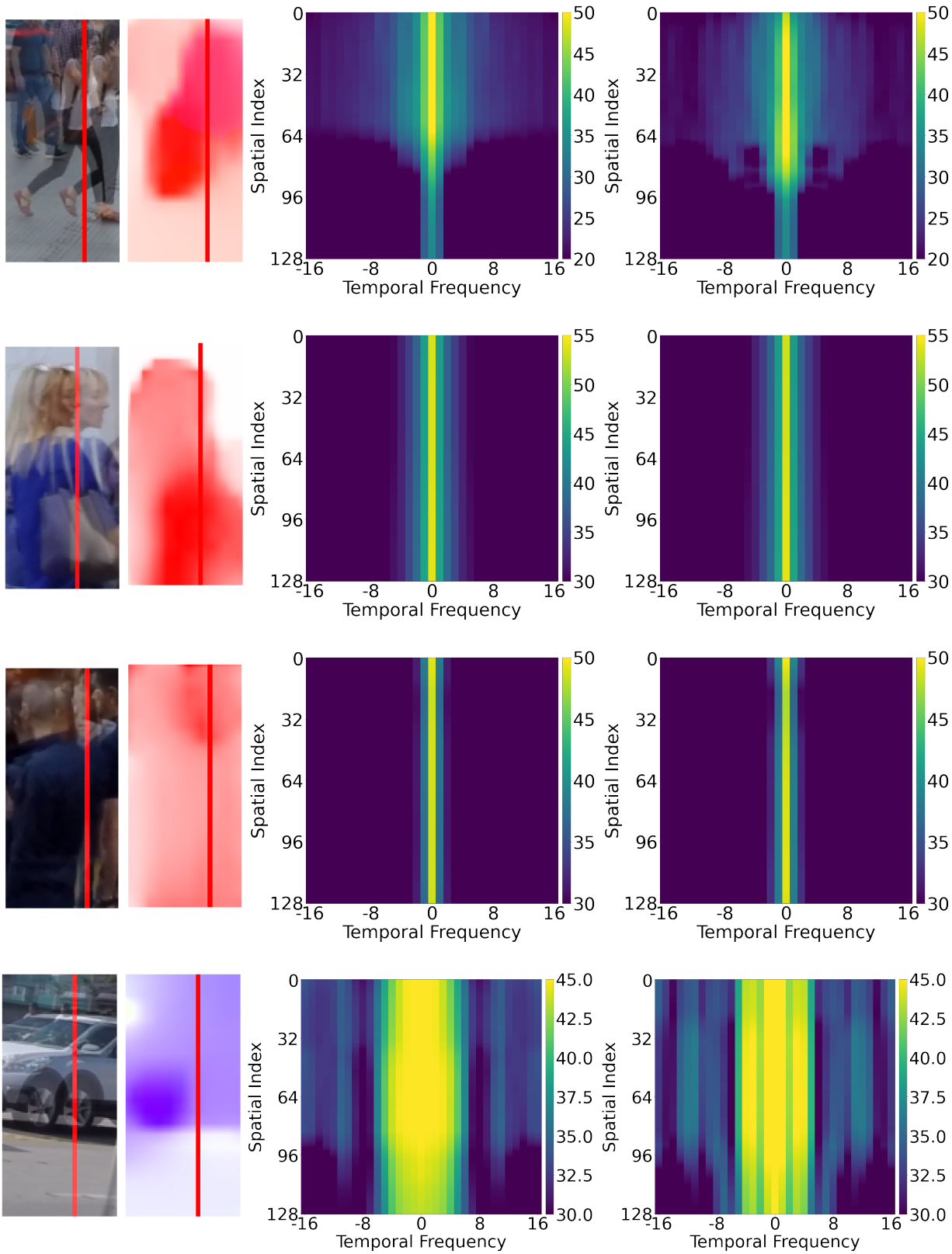

Figure A2: Fourier analysis of forward and backward motion. The first column shows the slice of pixels whose forward/backward motion are analyzed. The second column is the forward optical flow map. The third column is the temporal signal spectra of the vertical component of the forward displacement vectors. The forth column is the temporal signal spectra of the vertical components of the backward displacement vectors. The spectra shown are magnitude responses. Forward and backward motion have similar frequency responses. This is because most video sequences have less and smaller vertical motion. See Section A3.

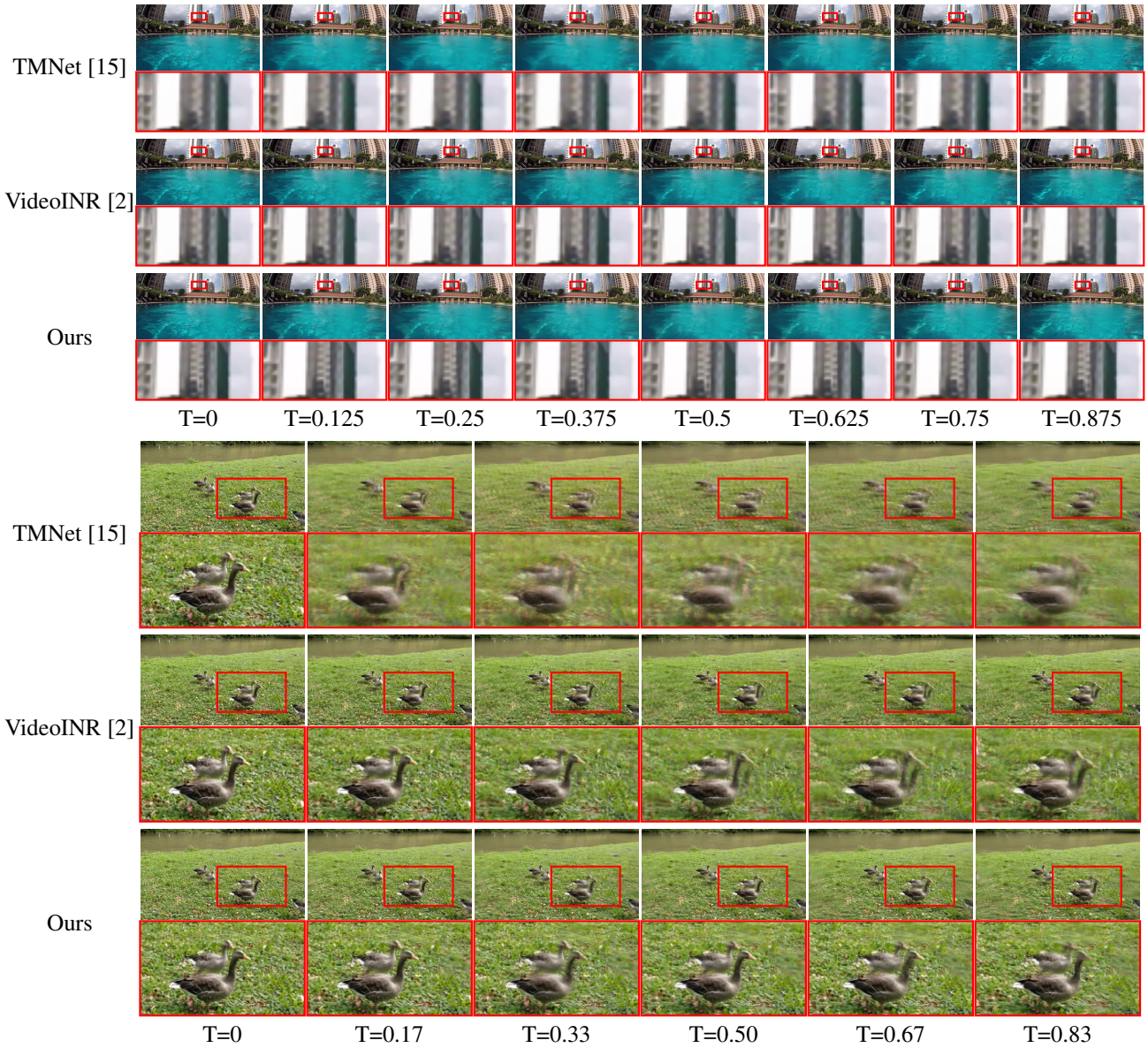

Figure A3: Subjective quality comparison. The temporal scaling factor of the upper example is 8 (in-distribution), and that of the lower example is 6 (out-of-distribution). Zoom in for better visualization. See Section A4.

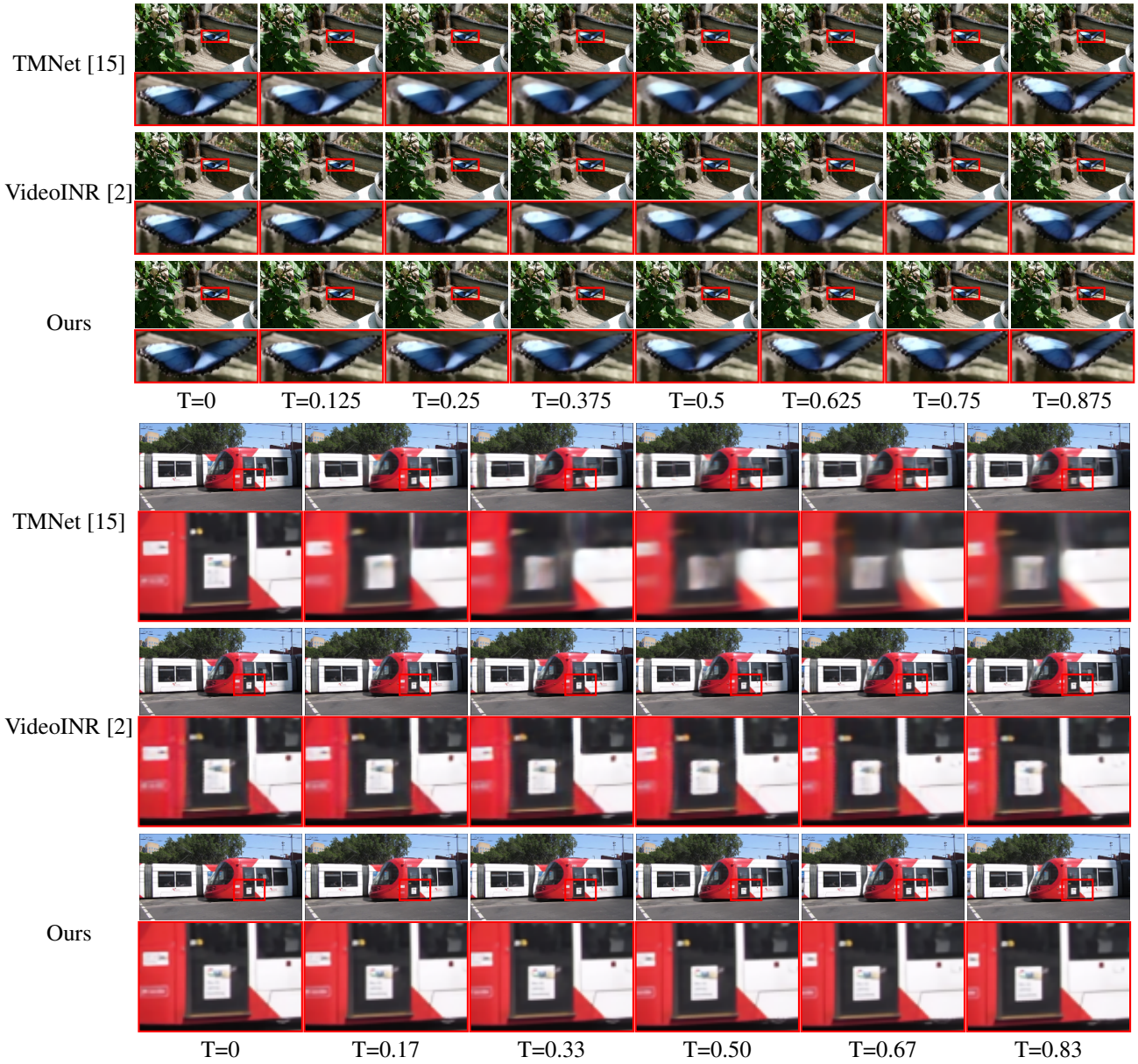

Figure A4: Subjective quality comparison. The temporal scaling factor of the upper example is 8 (in-distribution), and that of the lower example is 6 (out-of-distribution). Zoom in for better visualization. See Section A4.

Figure A5: Subjective quality comparison with different spatial scaling factors. We display the middle frame at $t = 0.5$. From left to right, the spatial scaling factors are 2, 4 (in-distribution) and 6 (out-of-distribution), respectively. Zoom in for better visualization. See Section A4.

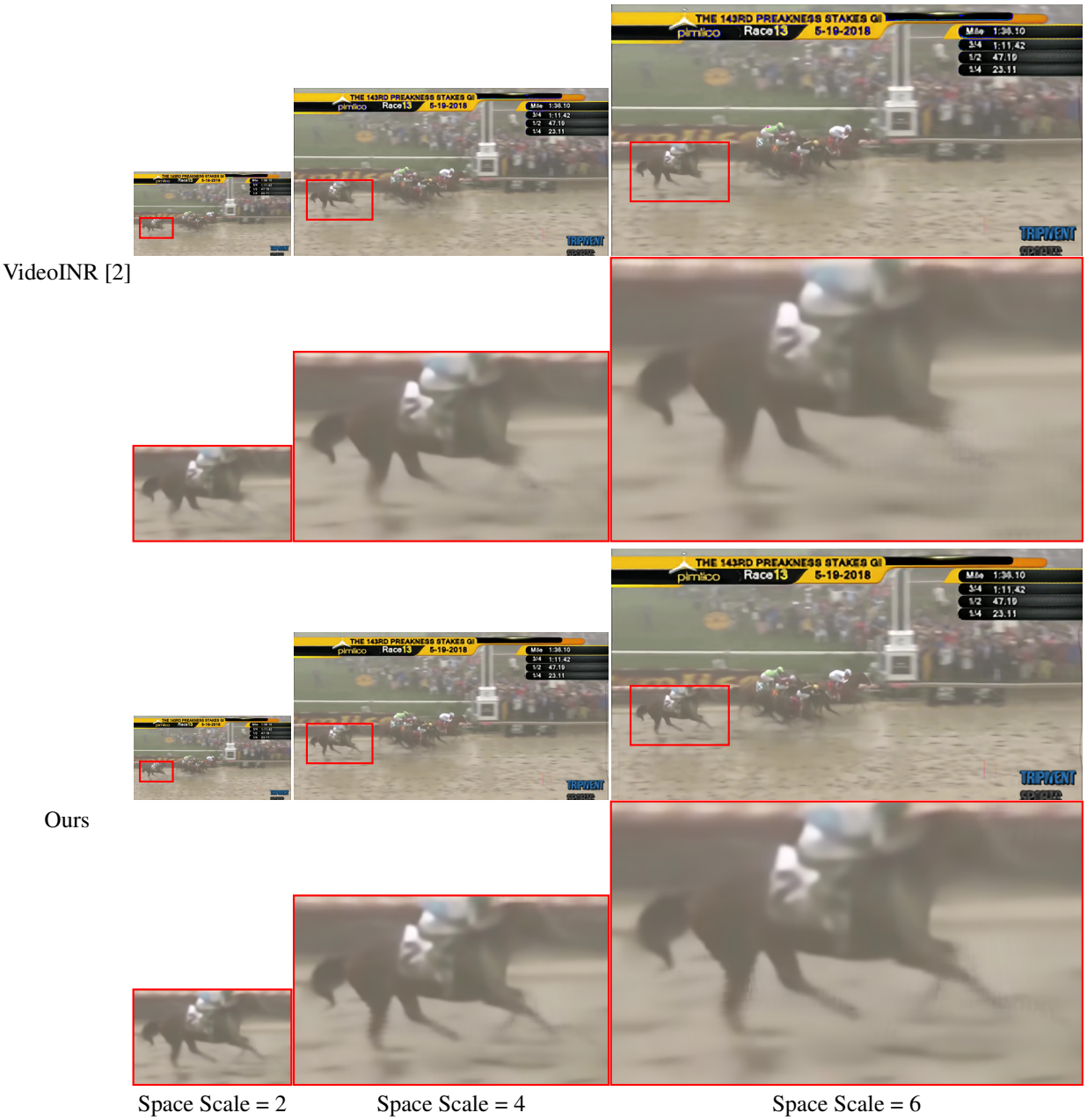

Figure A6: Subjective quality comparison with different spatial scaling factors. We display the middle frame at $t = 0.5$. From left to right, the spatial scaling factors are 2, 4 (in-distribution) and 6 (out-of-distribution), respectively. Zoom in for better visualization. See Section A4.

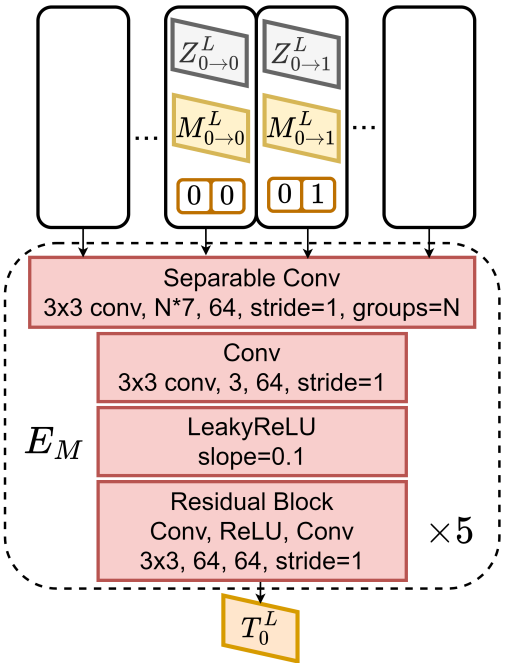

Figure A7: The network architecture of our motion encoder E_M. N is the number of motion samples we use. See Section A5.

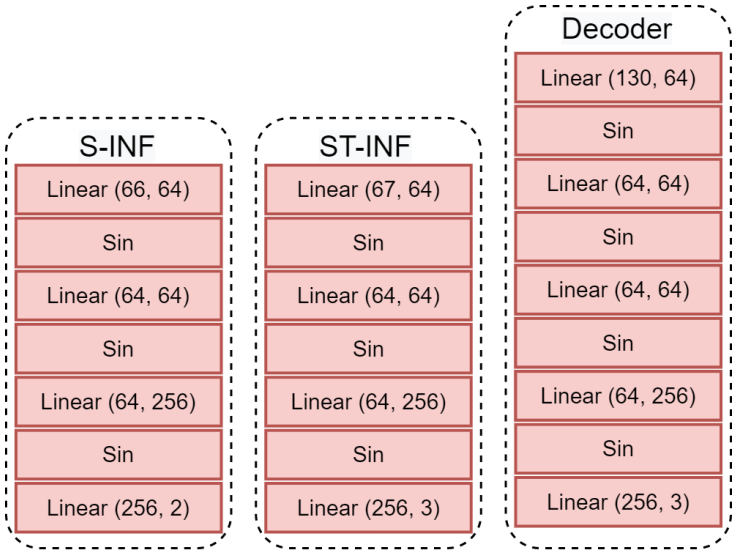

Figure A8: Shown from left to right are the network architectures of our S-INF, ST-INF and decoder, respectively.