EvEnhancer: Empowering Effectiveness, Efficiency and Generalizability for Continuous Space-Time Video Super-Resolution with Events

Supplementary Material

A. Overview

In this supplementary material, we first elaborate on the network architecture of the proposed EvEnhancer (Sec. B). Then, more ablation studies are conducted including 1) the multi-scale alignment in event-modulated alignment (EMA), 2) the bidirectional recurrence in bidirectional recurrent compensation (BRC), 3) the attention mechanism and positional encoding in local implicit video transformer (LIVT), and 4) the hyper-parameter setting of our models (Sec. C). Next, we investigate the temporal consistency of reconstructed frames by our EvEnhancer and compare it with existing state-of-the-art methods (Sec. D). Finally, more visual results on synthetic and real-world datasets are provided (Sec. E).

B. Network Architecture

In EvEnhancer, there are three parts: 1) feature extraction, 2) event-adapted synthesis module (EASM), and 3) local implicit video transformer (LIVT). Specifically, in the first step, for the LR frames, we use one 5×5 convolutional layer with LeakyReLU (LReLU) and 5 residual blocks to extract RGB features. As for voxelized event segments, we use the 3×3 convolution and another 5 residual blocks. The EASM contains two parts: 1) event-modulated alignment (EMA) and 2) bidirectional recurrent compensation (BRC) propagates the event stream across time and fuses it with acquired features in both directions to maximize the gathering of temporal information.

In EMA, all the convolutional layers are with the kernel size of 3×3 activated by LReLU. We use the deformable convolutional network [14] for feature alignment. In BRC, besides the first 5×5 convolution for event feature extraction and 1×1 convolutions in the channel attention mechanisms and feed-forward network [11], all the convolutional layers are also with the kernel size of 3×3. In LIVT, we use one $3 \times 3 \times 3$ convolutional layer to extract the spatiotemporal features. Then, another three 3D convolutions with the same kernel size are used to get the key K, query Q, and value V, where the temporal selection scheme is illustrated in Figure i. Finally, we use a 5-layer MLP with $[256, 256, 256, 256, 3]$ dimensions and GELU activations to produce the HR and HFR frames. In this work, we implement two model variants EvEnhancer and EvEnhancer-light. For the former, we set the number of event segments $M = 7$ and all the convolutional layers have 64 channels. For EvEnhancer-light, we reduce the event segment M to 5 and the channel number of the convolutional layers in LIVT to 16.

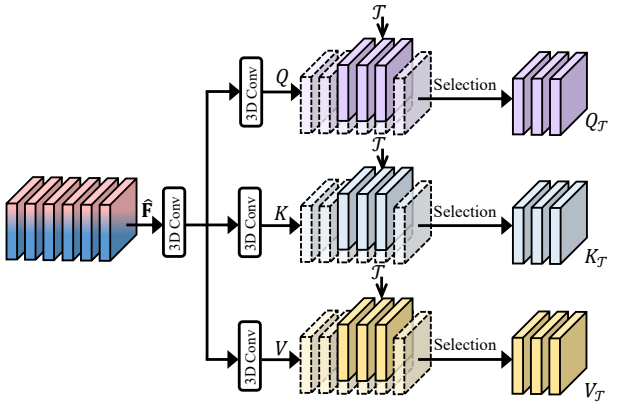

Figure i. The detail process of obtaining the key $K_{\mathcal{T}}$, query $Q_{\mathcal{T}}$, and value $V_{\mathcal{T}}$ near the target timestamp $\mathcal{T}$ in local implicit video transformer (LIVT).

Table i. Ablation studies on the event-modulated alignment (EMA). Metrics: PSNR (dB) / SSIM.

Alignment	In-Dist.	OOD
1× scale	33.21 / 0.9266	29.46 / 0.8560
1/2× scale	32.98 / 0.9231	29.32 / 0.8520
1/4× scale	32.92 / 0.9223	29.26 / 0.8502
multi-scale	**33.30 / 0.9279**	**29.54 / 0.8579**

The cosine positional encoding in LIVT is an extension of [1], formulated by

$$g(\delta\mathbf{C}) = [\sin(2^0\delta\mathbf{C}), \cos(2^0\delta\mathbf{C}), ..., \sin(2^{L-1}\delta\mathbf{C}), \cos(2^{L-1}\delta\mathbf{C})], \quad \text{(i)}$$

where $\delta\mathbf{C} = \{(\delta\tau, \delta x, \delta y)\}$ is the spatiotemporal relative coordinates and L is a hyper-parameter set to 10.

C. More Ablation Studies

Effect of the Multi-Scale Alignment in EMA. Table i shows that $1\times$ scale alignment plays a more critical role. However, the multi-scale manner ($1\times$, $\frac{1}{2}\times$, $\frac{1}{4}\times$) captures richer motion cues compared to the single-scale setting, leading to improved performance.

Effect of the Bidirectional Recurrence in BRC. Table ii presents the ablations of BRC, which involves bidirectional (forward and backward) recurrence in BRC. We also investigate the influence of channel attention mechanisms in BRC. As we can see, the model using only the forward or backward compensation shows the worst performance. When we incorporate attention in each direction, the performance increases. Moreover, by implementing bidirectional

Table ii. Ablation studies on the bidirectional recurrent compensation (BRC). Metrics: PSNR (dB) / SSIM.

Recurrent	Channel Attention	In-Dist.	OOD
fwd.		32.57 / 0.9176	28.96 / 0.8429
bwd.		32.55 / 0.9171	28.94 / 0.8424
fwd.	✓	32.79 / 0.9211	29.12 / 0.8467
bwd.	✓	32.79 / 0.9207	29.11 / 0.8467
fwd. & bwd.	✓	**33.30 / 0.9279**	**29.54 / 0.8579**

Table iii. Ablation studies on the attention mechanism and positional encoding in local implicit video transformer (LIVT). Metrics: PSNR (dB) / SSIM.

Attention Mechanism	Positional Encoding	In-Dist.	OOD
Neighborhood [4]	Cosine	33.07 / 0.9248	29.24 / 0.8492
Cross-scale	Learnable [8]	31.92 / 0.9086	28.37 / 0.8294
Cross-scale	Cosine	**33.30 / 0.9279**	**29.54 / 0.8579**

compensation with attention both forward and backward, the model performs the best, which achieves significant improvements.

Impact of the Cross-Scale Attention & Cosine Positional Encoding in LIVT. In LIVT, the query is obtained from the large-scale features at the HR grid after trilinear upsampling, while the key and value are obtained from the small-scale at the local LR grids nearest to the query. Therefore, we call it "cross-scale", which can capture spatiotemporal dependencies across LR and HR scales. Our cross-scale attention can be seen as a 3D cross-scale derivation of neighborhood attention [4]. As illustrated in Table iii, it exhibits suboptimal performance if we use this neighborhood attention directly. Besides, we encode and reshape the spatiotemporal relative coordinates $(\delta\tau, \delta x, \delta y) \in (-1, 1)$ from each query point to all pixel points within its local grid via the cosine positional encoding (Eq. i). Here, we investigate its impact by comparing it with another learnable positional encoding scheme [8]. Table iii shows that the model with cosine positional encoding is superior to the learnable one.

Hyper-parameter Setting. Here, we investigate the impact of event segments M in EASM, local grid size in LIVT, and the channel of learning video INR. As shown in Table iv, selecting an insufficient number of event segments, INR channels, or local grid size can lead to a degradation in reconstruction quality. Conversely, an over-large number of each can significantly decrease the model efficiency. Considering the balance between the performance and complexity, we implement the EvEnhancer-light and EvEnhancer using the settings in the last two rows as our baseline models to compare with other methods in this work.

D. Temporal Consistency

In Figure ii, we visualize the temporal profiles of VideoINR [3], MoTIF [2], and our EvEnhancer on the GoPro dataset [9]. We can observe that, the results of

Table iv. Ablation studies on hyperparameters including the number of event segments M, the number of INR channels, and local grid size of $T^G \times H^G \times W^G$. Metrics: PSNR (dB) / TFLOPs.

M	Local Grid	INR Channel	In-dist.	OOD	Params (M)
5	3×3×3	64	32.32 / 6.742	28.73 / 8.512	6.548
9	3×3×3	64	33.08 / 7.516	29.46 / 8.852	6.548
7	1×3×3	64	33.15 / 4.609	27.00 / 5.077	6.253
7	5×3×3	64	33.35 / 9.649	29.60 / 12.29	6.843
7	3×1×1	64	32.99 / 3.769	29.29 / 3.875	6.155
7	3×5×5	64	33.28 / 13.85	29.52 / 18.30	7.335
7	3×3×3	16	33.01 / 4.003	29.12 / 4.416	5.810
5	3×3×3	16	32.73 / 3.663	28.94 / 4.266	5.810
7	3×3×3	64	33.30 / 7.129	29.54 / 8.682	6.548

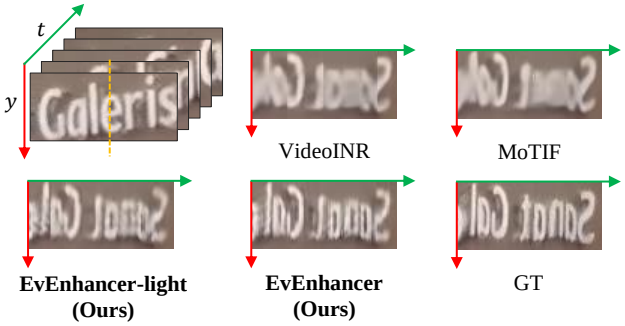

Figure ii. Comparison of temporal profile on the GoPro dataset [9] ($t = 12, s = 6$). We select a column (orange dotted lines) and observe the changes across time.

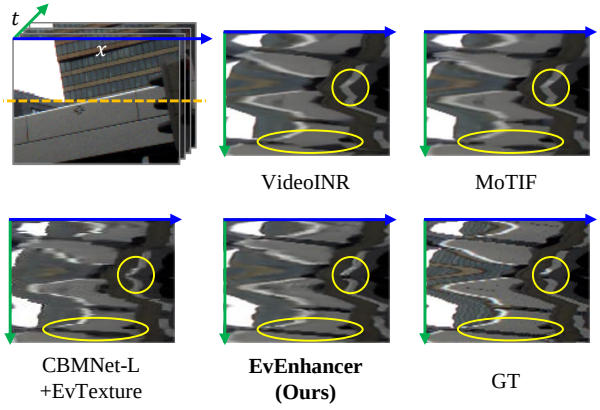

Figure iii. Comparison of temporal profile on the BS-ERGB dataset [13] ($t = 4, s = 4$). We select a row (orange dotted lines) and observe the changes across time.

VideoINR and MoTIF contain obvious noise, blurs, and heavy flickering artifacts, which indicates their poor temporal consistencies. In contrast, the profiles of EvEnhancer-light can guarantee better consistency but still contain discontinuity and artifacts. Our full model EvEnhancer shows more pleasant and smoother temporal profiles. Also, we visualize the temporal profiles on the real-world BS-ERGB dataset [13] in Figure iii, including event-based VFI + VSR methods (CBMNet-L [7] + EvTexture [6]). Our EvEnhancer maintains the best consistency.

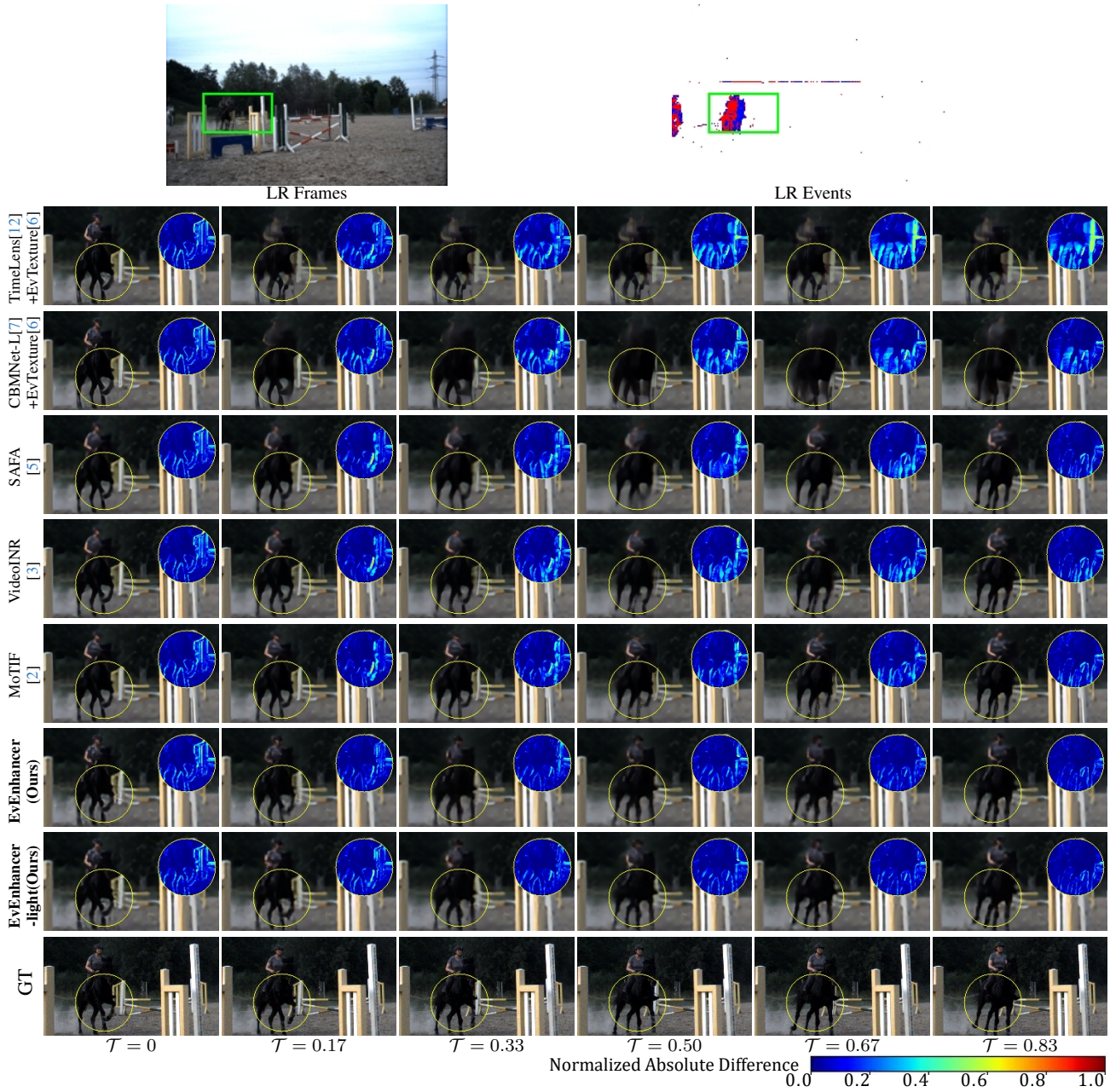

Figure iv. Qualitative comparison for OOD scale ($t = 6, s = 4$) on the BS-ERGB dataset [13]. We compare the normalized absolute difference maps (yellow boxes) for the same regions in each frame as in the GT frames.

E. More Visualization Results

Limited by the space of the main manuscript, in the supplementary material, we provide more qualitative results on both synthetic Adobe240 [10] and GoPro [9] datasets and real-world BS-ERGB dataset [13]. In Figure iv, we conduct comparisons on the real-world BS-ERGB dataset, where the spatiotemporal scales are OOD. We also calculate the difference maps between the reconstructed frames and the ground-truth (GT) to reveal the capacity for detail recovery. As we can see, our method can produce more preferable HR frames at any time. In Figure v, vi, and vii, there are more results for In-dist. and OOD scales on the BS-ERGB [13], Adobe240 [10], and GoPro [9] datasets, where the results demonstrate the superior effectiveness of our models. In Figure viii, we fix the temporal scale between two input LR frames as 6 and perform arbitrary spatial VSR. As seen in the reconstruction performance of the center frame at timestamp $\mathcal{T} = 0.5$, both our EvEnhancer-light and EvEnhancer can recover more textures.

References

[1] Hao-Wei Chen, Yu-Syuan Xu, Min-Fong Hong, Yi-Min Tsai, Hsien-Kai Kuo, and Chun-Yi Lee. Cascaded local implicit transformer for arbitrary-scale super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18257–18267, 2023. 1

[2] Yi-Hsin Chen, Si-Cun Chen, Yen-Yu Lin, and Wen-Hsiao Peng. Motif: Learning motion trajectories with local implicit neural functions for continuous space-time video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23131–23141, 2023. 2, 3, 5, 6, 7, 8

[3] 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*, pages 2047–2057, 2022. 2, 3, 5, 6, 7, 8

[4] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6185–6194, 2023. 2

[5] Zhewei Huang, Ailin Huang, Xiaotao Hu, Chen Hu, Jun Xu, and Shuchang Zhou. Scale-adaptive feature aggregation for efficient space-time video super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4228–4239, 2024. 3

[6] Dachun Kai, Jiayao Lu, Yueyi Zhang, and Xiaoyan Sun. EvTexture: Event-driven Texture Enhancement for Video Super-Resolution. In *Proceedings of the 41st International Conference on Machine Learning*, pages 22817–22839. PMLR, 2024. 2, 3

[7] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18032–18042, 2023. 2, 3

[8] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[9] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 2, 3, 7, 8

[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*, pages 1279–1288, 2017. 3, 6

[11] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023. 1

[12] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16155–16164, 2021. 3

[13] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 2, 3, 5

[14] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 1

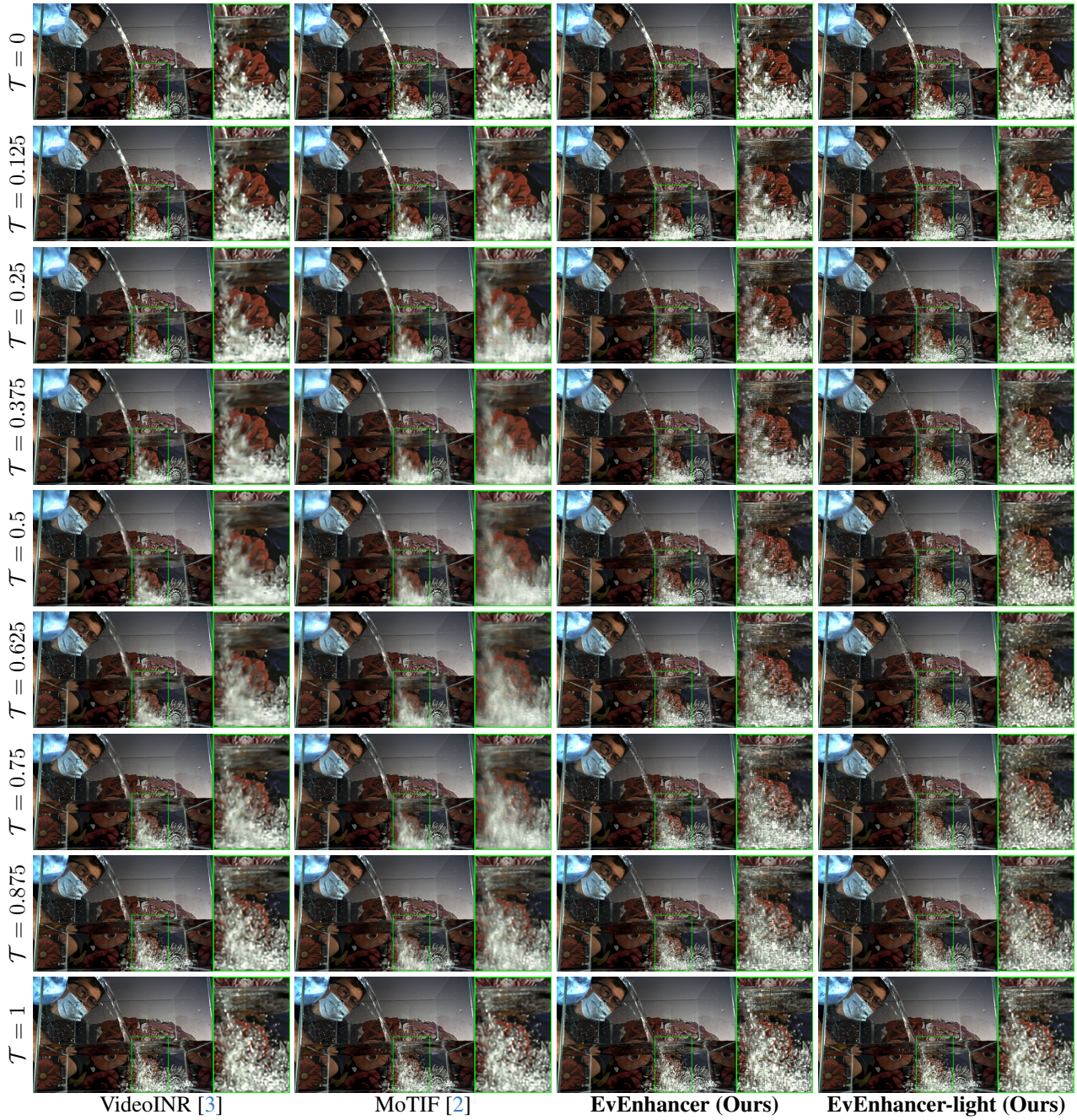

Figure v. Qualitative comparison for In-Dist. scale ($t = 8, s = 4$) on the BS-ERGB dataset [13]. Best zoom in for better visualization.

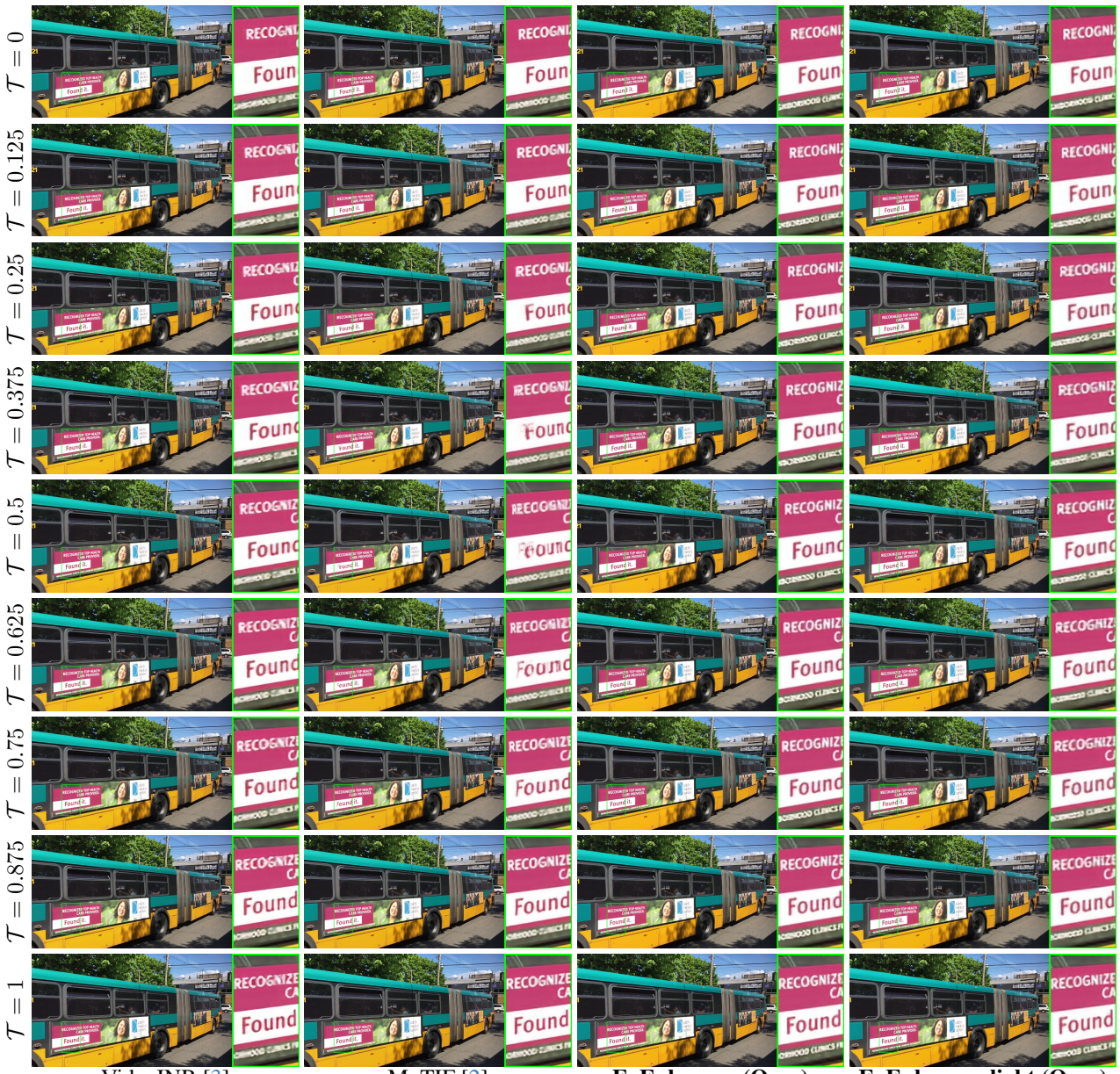

Figure vi. Qualitative comparison for In-Dist. scale ($t = 8, s = 4$) on the Adobe240 dataset [10]. Best zoom in for better visualization.

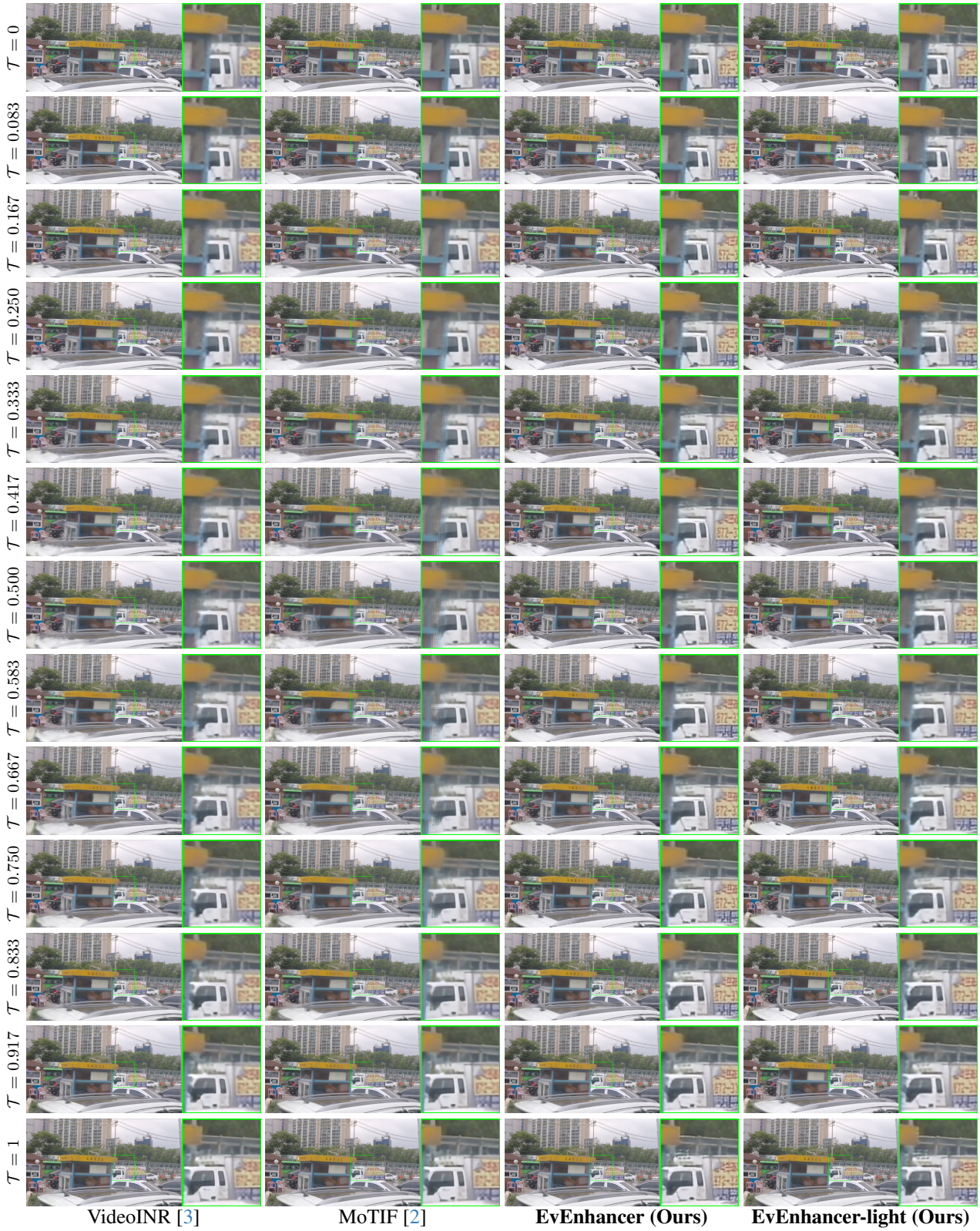

Figure vii. Qualitative comparison for OOD scale ($t = 12, s = 6$) on the GoPro dataset [9]. Best zoom in for better visualization.

Figure viii. Qualitative comparison for different spatial scale ($s = 2, 4, 6$), and fixed temporal scale $t = 6$ on the GoPro dataset [9]. We display the center frame at $\mathcal{T} = 0.5$. Best zoom in for better visualization.