

# AE2VID: Event-based Video Reconstruction via Aperture Modulation

## Supplementary Material

Chenxu Bai<sup>1,2\*</sup> Boyu Li<sup>1,2\*</sup> Peiqi Duan<sup>1,2†</sup> Xinyu Zhou<sup>3</sup> Hanyue Lou<sup>1,2</sup> Boxin Shi<sup>1,2,4†</sup>

<sup>1</sup> State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

<sup>2</sup> National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

<sup>3</sup> State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University

<sup>4</sup> PKU-AI<sup>2</sup> Robotics Joint Lab of Embodied AI

chenxu.bai@stu.pku.edu.cn {liboyu, duanqi0001, zhouxinyu, hylz, shiboxin}@pku.edu.cn

Our supplementary material is organized as follows: We first give discussions on the aperture shutter in Sec. 6 and more implementation details in Sec. 7. Secondly, we conduct several ablation studies in Sec. 8. Then, we compare computational efficiency with others in Sec. 9. Furthermore, lens parameters for real data capture are discussed in Sec. 10. Finally, more qualitative results are illustrated in Sec. 11.

We also provide a video (AE2VID\_supp\_video.mp4), which includes an animated illustration of AE2VID framework and video results on AMED and EvAid [3] datasets.

### 6. Discussions on the aperture shutter

The principle of aperture-modulation-triggered events has been formulated in Sec. 3.1 of main paper. In our implementation, motorized aperture shutters are adopted for modulation due to their stability, but in practice, we find that manual adjustment of common C-mount lenses can achieve similar effects. Besides, there are also other choices of Transmittance Adjustment Devices for aperture modulation, such as rotary polarization reducers or liquid crystal optical switches [1], but their shading properties are inferior.

A sample of the aperture-modulation-triggered event stream with the aperture shutter is shown in Fig. 5 (a)-(c), and the corresponding frame reconstructed by AENet is shown in Fig. 5 (d). In our experiments, we observe that the motorized aperture shutter exhibits an asymmetric opening process. Specifically, we capture a uniformly illuminated whiteboard using the aperture shutter and manual rotation, respectively, and obtain the normalized FPE temporal matrices shown in Fig. 5 (e)&(f). As can be seen, the right side of the FPE temporal matrix for the aperture shutter is generally smaller than the left side, indicating that the right side is triggered earlier than the left side under the same illumination. In contrast, manual adjustment yields a more uniform distribution. We further derive a drift matrix from these matrices. Although this has only a minor impact on reconstruction, since its magnitude is negligible ( $\sim 1\%$ ) relative to timestamps, we nevertheless correct all real-data

results using the drift matrix.

### 7. More implementation details

Our modulation strategy consists of several observation windows, each with an equivalent length  $\tau$ . The observation window can be further divided into three stages: the aperture opening process, the interval where the aperture is on, and the aperture closing process. Note that there are no intervals between observation windows, which means we will reopen the aperture immediately after the closing process. Among them, the opening and closing process both takes  $\delta t$ , and the interval takes  $\tau - 2\delta t$ . The detailed discussions of these parameters are in Sec. 10.

For the aperture opening process, due to the high temporal resolution of event cameras, we can record the static background information of the scene in a short period of  $\delta t$

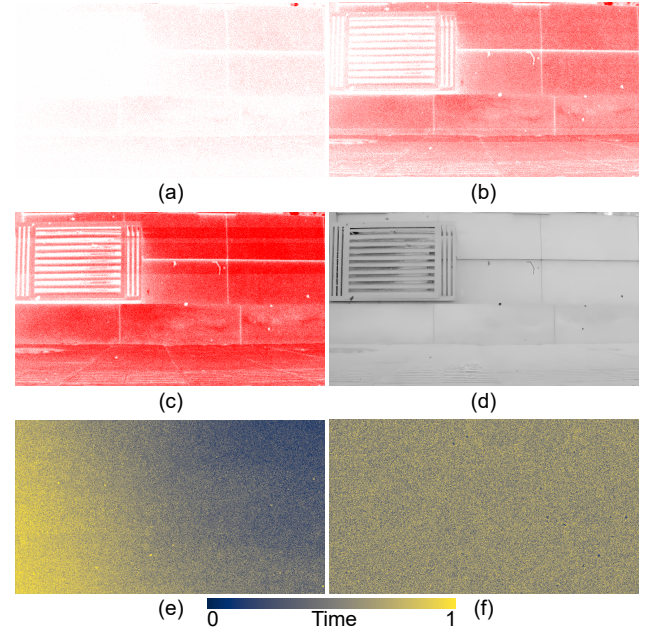


Figure 5. (a)-(c) Visualization of aperture-modulation-triggered events in chronological order. (d) Reconstructed frame from aperture-modulation-triggered events. (e) FPE temporal matrix of aperture shutter. (f) FPE temporal matrix of manual rotation.

\*Equal contribution.

†Corresponding authors.

without much loss of motion cues. As  $\delta t$  is relatively much smaller than  $\tau$ , we reconstruct one frame  $\hat{\mathbb{I}}_i^A$  from each opening process using our proposed AENet. AENet is composed of FIR, IDN, and HSG modules, where we choose SwinIR [6] as the IDN module for denoising.

For the interval where the aperture is on, motion-triggered events are exploited to reconstruct a continuous video sequence. The raw events are first converted into voxel grids with  $b = 5$  bins and subsequently processed by MENet. MENet is composed of recurrent blocks  $\mathcal{R}$  unrolled for  $K$  time steps. Each block  $\mathcal{R}$  comprises bidirectional LSTM [11] modules and a mixer  $\mathcal{M}$ . The mixer  $\mathcal{M}$  takes as input the forward and backward LSTM predictions,  $\hat{\mathbb{I}}_{i,k}^{M,\text{fwd}}$  and  $\hat{\mathbb{I}}_{i,k}^{M,\text{bwd}}$ , together with the reference frames reconstructed by AENet,  $\hat{\mathbb{I}}_i^A$  and  $\hat{\mathbb{I}}_{i+1}^A$ , as well as the relative timestamp  $k$ , and produces a pixel-wise weight map  $\alpha_{i,k}$ . This weight map is then applied to the candidates to yield the final predictions.

For the aperture closing process, the captured events contain insufficient useful information and are therefore discarded, resulting in missing frames within the time  $\delta t$ . However, since we have already reconstructed the frames immediately preceding the closure and those with subsequent aperture opening, we employ the RIFE model [5] to interpolate the frames corresponding to this gap, thereby restoring a temporally continuous video sequence.

For all the compared methods, we use their officially released pretrained checkpoints. Specifically, for E2VID [8, 9], we use the E2VID\_lightweight checkpoint; for FireNet [10], we use the firenet\_1000 checkpoint; for SPADE-E2VID [2], we use the SPADE\_E2VID checkpoint; for V2V-E2VID [7], we use the v2v\_e2vid\_10k checkpoint; for others, we use their respective checkpoints.

## 8. Ablation studies

To verify the effectiveness of each component in our framework, we conduct several ablation studies on the EvAid [3] dataset and show the results in Table 3. Firstly, we show the advantage of a bidirectional pipeline compared with a unidirectional pipeline trained with the same setting (denoted as “Unidirectional”). Secondly, to validate the effectiveness of reference frames  $\hat{\mathbb{I}}_i^A$  and  $\hat{\mathbb{I}}_{i+1}^A$ , we change the input to the pixel-wise mixer  $\mathcal{M}$  to simply two candidates  $\hat{\mathbb{I}}_{i,k}^{M,\text{fwd}}$ ,  $\hat{\mathbb{I}}_{i,k}^{M,\text{bwd}}$  and the event voxel  $V_{i,k}^M$  (denoted as “Mix-2”). Furthermore, we test the validity of our two-stage training scheme by directly training the whole pipeline for 20 epochs (denoted as “Train-whole”). Besides, we conduct ablation studies on our loss functions. We verify the effectiveness the pseudo frame  $\hat{\mathbb{I}}^{A'}$  output by the HSG module by excluding the  $\ell_1$  loss  $\|\hat{\mathbb{I}}^{A'} - \hat{\mathbb{I}}^A\|_1$  from our loss function (denoted as “w/o  $\hat{\mathbb{I}}^{A'}$ ”). Additionally, we conduct experiments on calculating the temporal consistency loss for the full sequence (denoted as “Full-TC”) and excluding this loss

Table 3. Ablation study results on EvAid [3].

Method	MSE↓	SSIM↑	MS-SSIM↑	LPIPS↓
Unidirectional	0.124	0.539	0.396	0.503
Mix-2	0.039	0.694	0.540	0.430
Train-whole	0.043	0.692	0.530	0.428
w/o $\hat{\mathbb{I}}^{A'}$	0.039	0.688	0.533	0.415
Full-TC	0.039	0.700	0.531	0.422
w/o TC	0.039	0.693	0.537	0.414
w/o FIR	0.041	<b>0.707</b>	0.531	0.418
w/o IDN	0.040	0.663	0.514	0.427
Conv-HSG	0.044	0.631	0.458	0.499
Ours	<b>0.037</b>	<b>0.707</b>	<b>0.544</b>	<b>0.411</b>

Table 4. Computation efficiency comparison results.

	Params	MACs	Time
E2VID [8, 9]	10.71 M	29.79 G	2.38 ms
FireNet [10]	37.78 K	2.46 G	1.52 ms
ETNet [13]	16.65 M	35.84 G	2.35 ms
SPADE-E2VID [2]	11.46 M	103.29 G	5.71 ms
PAEVSNN [14]	4.53 M	132.54 G	13.64 ms
BDE2VID [4]	19.35 M	54.45 G	7.16 ms
V2V-E2VID [7]	10.71 M	29.79 G	2.38 ms
Ours (w/o IDN)	53.08 M	73.53 G	4.68 ms

(denoted as “w/o TC”). Finally, the design of AENet is verified by the ablation of three components: ablation of FIR (“w/o FIR”) by training IDN to learn frames from aperture events, ablation of IDN (“w/o IDN”) by removing it, and ablation of HSG by replacing it with a convolution layer (“Conv-HSG”). From the comparison, we can observe that all the alternative models have degraded performance, while ours achieves the best.

## 9. Computational efficiency

We compare the number of parameters (Params), Multiply-Accumulate Operations (MACs), and inference time in Table 4. For fair comparison, all methods are tested on a single NVIDIA GeForce RTX 4090 GPU and an Intel i7-13700K CPU. The input resolution is  $256 \times 256$ , and we average the results over 100 frames. Note that as the IDN module can easily be replaced with other denoising networks and pre-computed offline, we do not include it in our framework in the statistics. It can be observed that ours achieve comparable MACs and inference time with previous methods.

## 10. Discussions on real data capture parameters

In this section, we are going to discuss the impact of real data capture parameters on the performance of our framework. There are three parameters mentioned in Sec. 4.1 of main paper, namely the final aperture position  $A_E$ , the aperture speed  $v_A$ , and the length of each observation window  $\tau$ .

To achieve precise control over scene illumination and motion dynamics, we employ a constant direct-current light

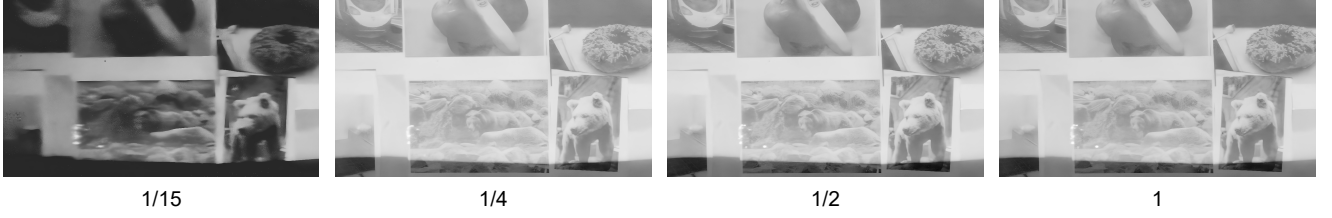


Figure 6. Reconstructed first frame comparisons with different final aperture positions. We compare 1/15, 1/4, 1/2, and the full size of the maximum aperture. Please zoom in for more details.



Figure 7. Reconstructed first frame comparisons with different aperture speeds. We compare 1/32, 1/8, 1/2, and the maximum speed. Please zoom in for more details.

source within an indoor setting and utilize a motorized rail for speed regulation, thereby ensuring environmental consistency with most capture scenarios. By systematically varying the camera parameters, we conduct the following qualitative analyses to identify the parameter configuration. Our capturing system includes a Prophesee EVK4 event camera with resolution  $1280 \times 720$  and a Computar LensConnect BH Series Variable Focal Length Lens<sup>1</sup>.

**Final aperture position.** We investigate the impact of the final aperture setting  $A_E$  by varying it across 1/15, 1/4, 1/2, and the full size of the maximum aperture. The qualitative comparisons of the first reconstructed frame with aperture-modulation-triggered events  $\hat{\mathbb{I}}_0^A$  are shown in Fig. 6. As illustrated, setting  $A_E$  to a mere 1/15 of the maximum aperture yields degraded reconstructions characterized by noise and blur, primarily stemming from incomplete event triggering and diffraction limits [1]. However, the method demonstrates robustness with negligible quality drop when  $A_E$  is set to 1/4, 1/2, or the full aperture. To minimize the loss of motion cues during the aperture transition, we empirically set  $A_E$  to 1/4 of the maximum aperture.

**Aperture speed.** We further compare configurations where the aperture speed  $v_A$  is set to 1/32, 1/8, 1/2, and the maximum speed. The first reconstructed frames  $\hat{\mathbb{I}}_0^A$  are shown in Fig. 7. It can be observed that when  $v_A$  is 1/32 of the maximum speed, the reconstructed frame is motion blurred on the left as the foreground object, *i.e.*, the colorchecker, has moved into the captured scene. In 1/8 and 1/2 scenarios,

the frame exhibits more pepper and salt noises compared to the full speed, perhaps because the slow opening process incurs more noise in event triggering. Therefore, we chose the full speed in our experiments.

**Interval.** The length of each observation window  $\tau$  is also a critical parameter for controlling the capture process. We evaluate the impact of varying interval  $\tau - 2\delta t$  across  $\{0, 1, 5, 10\}$  seconds, as visualized in Fig. 8. Please note that as different intervals exhibit problems at different timestamps (frame indices), each row corresponds to a different timestamp (frame index). Qualitative analysis reveals that a continuous opening-and-closing scheme (where  $\tau - 2\delta t = 0$ ) yields pronounced interpolation artifacts due to the severe loss of motion cues (see row 1, left). Similarly, a short interval of  $\tau - 2\delta t = 1s$  exhibits persistent motion artifacts, as evidenced by the colorchecker (see row 2, left). Conversely, an excessively prolonged interval of  $\tau - 2\delta t = 10s$  results in the degradation of background details during reconstruction (see row 3, left). Consequently, we empirically adopt  $\tau - 2\delta t = 5s$  to achieve a trade-off between motion fidelity and background preservation.

## 11. More qualitative results

More qualitative results on EvAid [3] are shown in Fig. 9, and more results on HQF [12] are in Fig. 10. Additionally, more results on our real-captured AMED dataset are illustrated in Fig. 11 and Fig. 12. From the comparison, we can see the superior detail-preserving and scene-reconstruction performance of our proposed method.

<sup>1</sup><https://www.edmundoptics.cn/p/9---50mm-lensconnect-bh-series-variable-focal-length-lens/53086/>

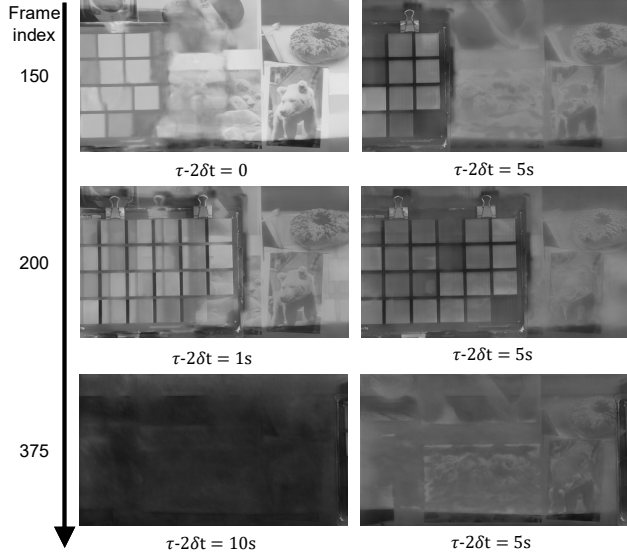


Figure 8. Reconstructed frame comparisons with different intervals. Note that each row corresponds to one different timestamp (frame index). We compare the results of  $\tau - 2\delta t = 5$  seconds with  $\tau - 2\delta t = 0, 1, 10$  seconds.

## References

- [1] Yuhan Bao, Lei Sun, Yuqin Ma, and Kaiwei Wang. Temporal-mapping photography for event cameras. In *Proc. of European Conference on Computer Vision (ECCV)*, 2024. [1](#), [3](#)
- [2] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. SPADE-E2VID: Spatially-adaptive denormalization for event-based video reconstruction. *IEEE Transactions on Image Processing*, 30:2488–2500, 2021. [2](#), [5](#), [6](#)
- [3] Peiqi Duan, Boyu Li, Yixin Yang, Hanyue Lou, Minggui Teng, Xinyu Zhou, Yi Ma, and Boxin Shi. EventAid: Benchmarking event-aided image/video enhancement algorithms with real-captured hybrid dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6959–6973, 2025. [1](#), [2](#), [3](#), [5](#)
- [4] Pinghai Gao, Longguang Wang, Sheng Ao, Ye Zhang, and Yulan Guo. Enhancing event-based video reconstruction with bidirectional temporal information. *IEEE Transactions on Multimedia*, 27:4831 – 4843, 2025. [2](#), [5](#), [6](#), [7](#), [8](#)
- [5] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proc. of European Conference on Computer Vision (ECCV)*, 2022. [2](#)
- [6] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021. [2](#)
- [7] Hanyue Lou, Jinxiu Liang, Minggui Teng, Yi Wang, and Boxin Shi. V2V: Scaling event-based vision through efficient video-to-voxel simulation. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2025. [2](#), [5](#), [6](#), [7](#), [8](#)
- [8] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-Video: Bringing modern computer vision to event cameras. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [2](#), [5](#), [6](#), [7](#), [8](#)
- [9] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1964–1980, 2021. [2](#), [5](#), [6](#), [7](#), [8](#)
- [10] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert E. Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proc. of Winter Conference on Applications of Computer Vision (WACV)*, 2020. [2](#), [5](#), [6](#), [7](#), [8](#)
- [11] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional LSTM network: a machine learning approach for precipitation now-casting. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2015. [2](#)
- [12] Timo Stofregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *Proc. of European Conference on Computer Vision (ECCV)*, 2020. [3](#), [6](#)
- [13] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proc. of International Conference on Computer Vision (ICCV)*, 2021. [2](#), [5](#), [6](#), [7](#), [8](#)
- [14] Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, and Yonghong Tian. Event-based video reconstruction via potential-assisted spiking neural network. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. [2](#), [5](#), [6](#), [7](#), [8](#)



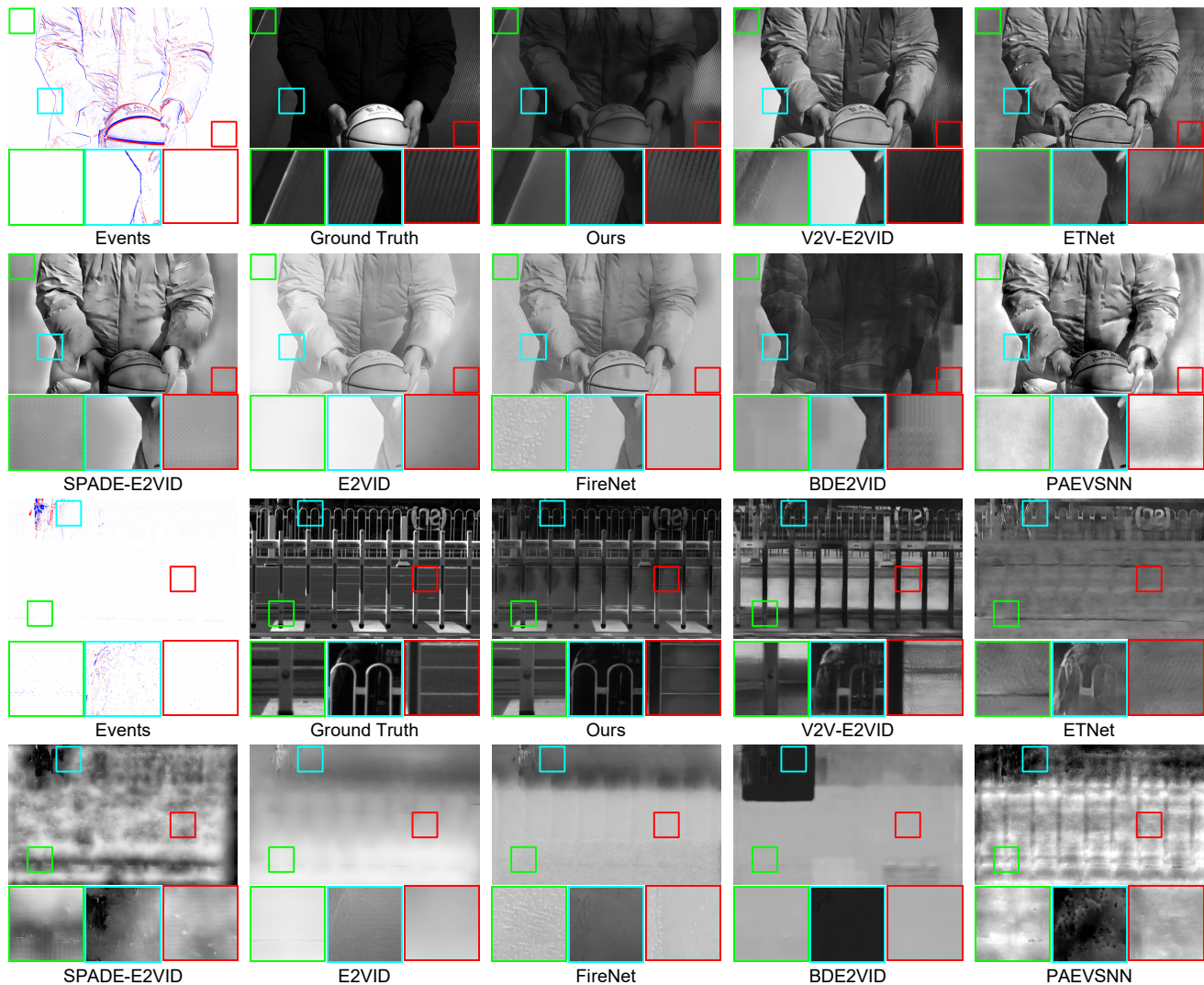


Figure 9. More qualitative experiment results on EvAid [3]. We compare with V2V-E2VID [7], ETNet [13], SPADE-E2VID [2], E2VID [8, 9], FireNet [10], BDE2VID [4], and PAEVSNN [14].

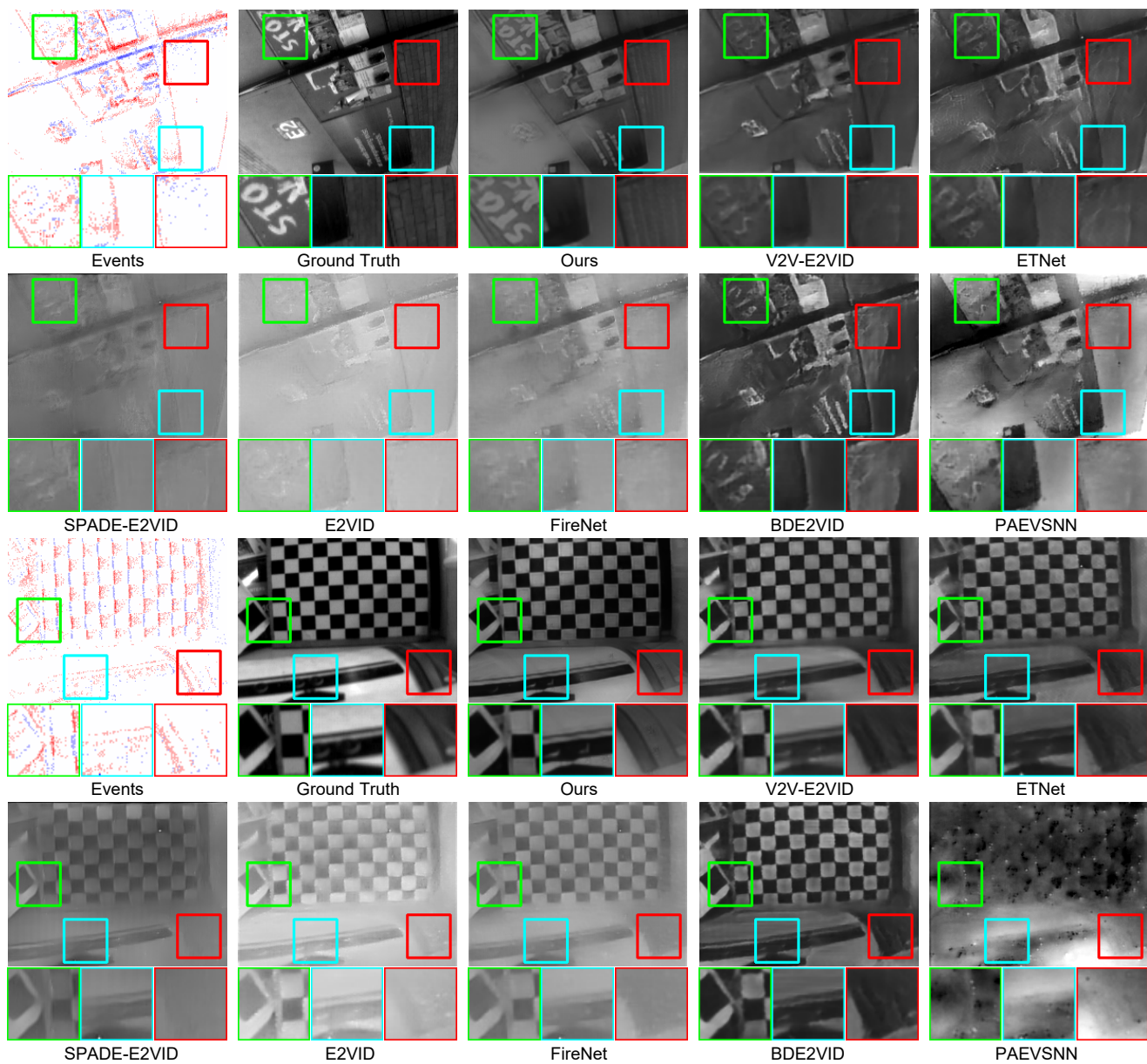


Figure 10. More qualitative experiment results on HQF [12]. We compare with V2V-E2VID [7], ETNet [13], SPADE-E2VID [2], E2VID [8, 9], FireNet [10], BDE2VID [4], and PAEVSNN [14].

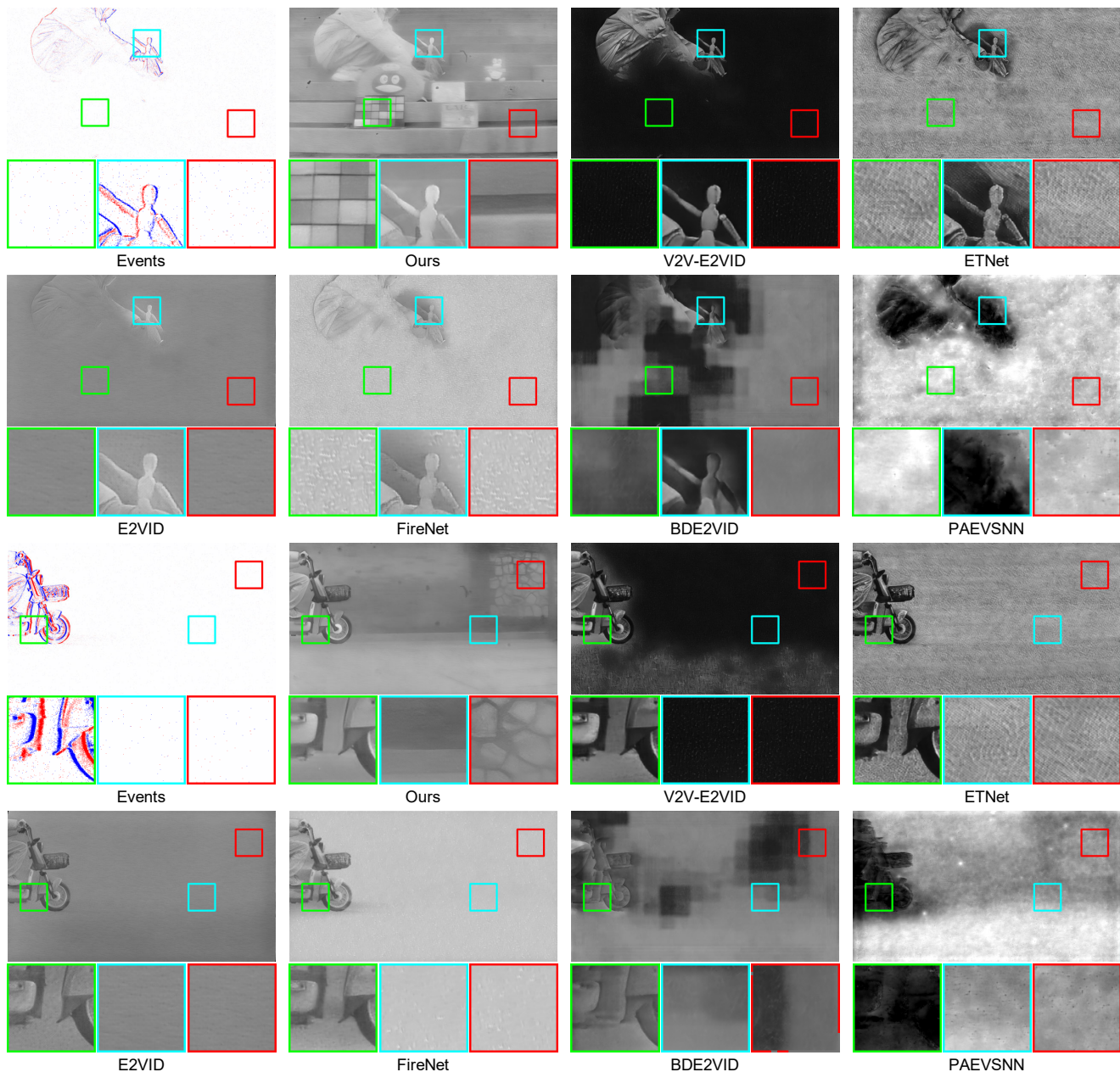


Figure 11. More qualitative experiment results on our real-captured AMED dataset (Part 1) with corresponding input motion-triggered events. We compare with V2V-E2VID [7], ETNet [13], E2VID [8, 9], FireNet [10], BDE2VID [4], and PAESNN [14].



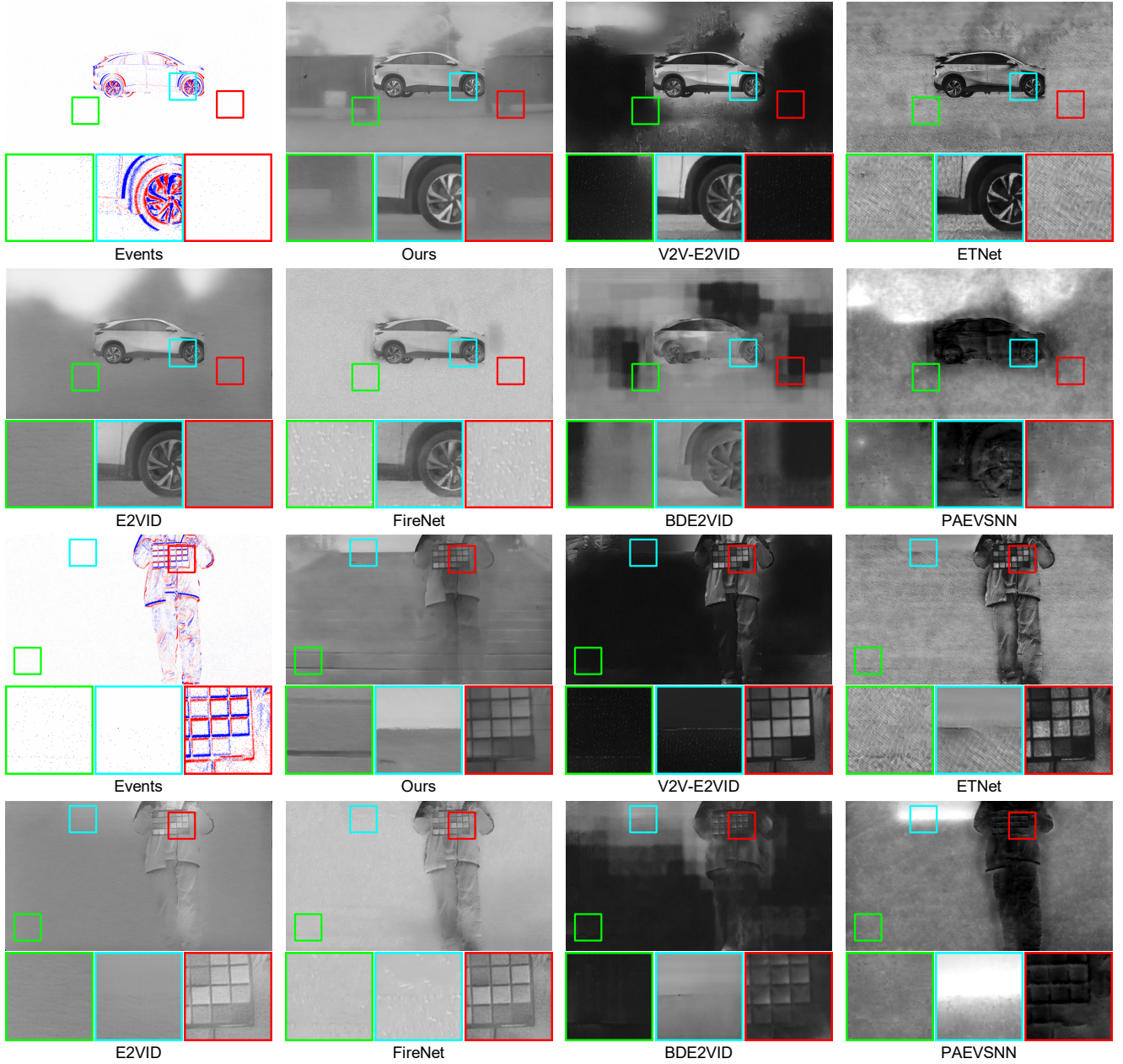


Figure 12. More qualitative experiment results on our real-captured AMED dataset (Part 2) with corresponding input motion-triggered events. We compare with V2V-E2VID [7], ETNet [13], E2VID [8, 9], FireNet [10], BDE2VID [4], and PAEVSNN [14].