AE2VID: Event-based Video Reconstruction via Aperture Modulation

Chenxu Bai[1,2*] Boyu Li[1,2*] Peiqi Duan[1,2†] Xinyu Zhou[3] Hanyue Lou[1,2] Boxin Shi[1,2,4†]

[1] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[3] State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University

[4] PKU-AI[2] Robotics Joint Lab of Embodied AI

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

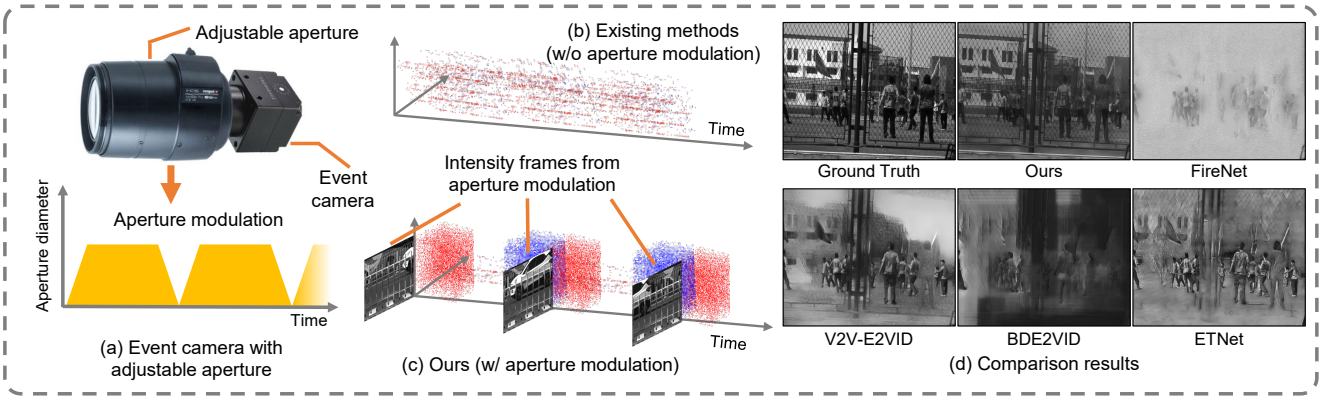

Figure 1. (a) We propose a framework that dynamically modulates the adjustable aperture of an event camera for video reconstruction. The aperture periodically opens and closes, injecting controlled illumination variations that encode scene information. (b) Existing methods operate without aperture modulation and thus rely solely on motion-triggered events, which are typically spatially sparse and provide little information about static regions. (c) Leveraging aperture modulation, our method reconstructs intermediate dense intensity frames, effectively complementing the sparse motion-triggered events to enhance fidelity. (d) The comparison results with FireNet [33], V2V-E2VID [26], BDE2VID [9], and ETNet [41]. Our method better preserves the static regions and overall scene structure than existing methods.

Abstract

Event-based video reconstruction seeks to recover high-speed, high-dynamic-range videos from event streams. While existing approaches rely exclusively on motion-triggered events, these events are inherently sparse and primarily capture dynamic regions. Therefore, they often suffer from error accumulation and degraded quality in regions with few events. In this work, we introduce aperture-modulation-triggered events as a complementary mechanism to enrich the captured scene information. Specifically, we periodically modulate the aperture to actively generate dense event signals, thereby encoding intensity cues even in static or low-motion regions. Building upon this idea, we design an AE2VID framework that jointly leverages aperture-modulation-triggered and motion-triggered events to enhance the fidelity of predictions. The proposed framework consists of two subnetworks for the dedicated processing of both event types. We further collect a real dataset and validate the effectiveness of our method. Extensive experiments show our superiority over state-of-the-art methods. Code and data will be available at https://github.com/a1henu/AE2VID/.

1. Introduction

Event cameras [4] are bio-inspired sensors that capture high-speed, high-dynamic-range scene information [30, 31]. Unlike spike cameras [16, 43], another type of neuromorphic sensor that records scenes through temporal integration, event cameras directly measure intensity changes. As a practical way to bridge event cameras with off-the-shelf frame-based algorithms, event-to-video reconstruction has attracted increasing attention [9, 11, 46]. The reconstructed

*Equal contribution.
†Corresponding authors.

videos can benefit many downstream computer vision tasks, such as 3D reconstruction [12], object recognition [42], and instance segmentation [24].

Early methods for event-based video reconstruction attempted to use hand-crafted operations [27, 32]. With the development of deep learning, recurrent learning-based methods [30, 31, 33, 45] have become the mainstream choice. Recently, some methods have incorporated diffusion-based priors [44, 49] to enhance the perceptual quality. However, as motion-triggered events are usually sparsely distributed along the edges (Fig. 1 (b)), this task is still highly ill-posed. Despite the improvements in model capacity, problems such as prediction error accumulation and static regions could restrict their performance (the building in the background and chain-link fence in the foreground of Fig. 1 (d)).

In order to tackle the above problems, additional dense observation of the scene radiance is necessary. An intuitive approach is to incorporate an additional frame-based camera as a supplementary input to the network [20, 38]. However, this may cause spatial-temporal misalignments [7] and incur additional costs. Alternatively, active lighting could be used to modulate the intensity received by cameras [13]. Although they can provide dense information of the scene by triggering events on almost every pixel, it is infeasible in outdoor or uncontrolled environments. Besides, static background noisy events can be jointly modeled to achieve reconstruction [10], but such an approach may encounter bottlenecks in outdoor scenes due to the presence of non-discriminative noisy events. To robustly acquire dense observations of the scene with fewer additional costs, we introduce aperture shutters [1, 36] for light intensity modulation.

The aperture is a ubiquitous and easily controllable component of most imaging systems. By adjusting the aperture, we can modulate the irradiance at each pixel (Fig. 1(a)), thereby triggering events for dense predictions that complement the sparsity of motion-triggered events. These predictions facilitate static background reconstruction and mitigate error accumulation, effectively addressing the aforementioned challenges (Fig. 1 (c)). However, the aperture-modulated video reconstruction still faces two challenges: (1) Effectively integrating aperture-modulation-triggered and motion-triggered events, which differ in spatial density, necessitates the design of novel architectures and specialized training strategies. (2) Acquiring real data under our modulation strategy, which involves dynamically closing and reopening the aperture, requires a dedicated capture system.

In this paper, we propose an event-based video reconstruction framework, which combines aperture-modulation-triggered and motion-triggered events for more robust predictions. We observe that events triggered by aperture modulation could provide dense scene priors to assist video reconstruction. Additionally, we design a reconstruction framework capable of adaptively fusing events from both sources to produce a continuous high-quality video. Furthermore, we design a modulation strategy consisting of periodically closing and reopening the aperture (Fig. 1(a)) to mitigate error accumulation. Finally, we employ our strategy on diverse scenes and construct a new real-world dataset for comprehensive evaluation. Overall, our contributions are:

- We are the first to introduce an aperture modulation strategy into event-based video reconstruction, which could solve slow initialization and error accumulation problems in mainstream methods based on motion-triggered events.
- We design a framework named AE2VID consisting of two dedicated subnetworks AENet and MENet, which could fuse the information from aperture-modulation-triggered and motion-triggered events in an efficient way, and output video frames with high speed and high dynamic range.
- We collect a real-world captured Aperture-modulation- and Motion-triggered Events Dataset (AMED), where we adopt our modulation strategy and specify several aperture control parameters. This dataset is utilized for evaluating the performance and inspiring future research.

Extensive experiments on both semi-real and real data show the superior background preservation ability and overall scene reconstruction quality of our approach.

2. Related work

2.1. Reconstruction with motion-triggered events

Early attempts for event-based video reconstruction employed extended Kalman filters through 2D Poisson integration [19]. Later, Bardow *et al*. [2] designed a variational energy minimization framework to simultaneously reconstruct video frames and optical flow. Scheerlinck *et al*. [32] proposed high-pass filters to process events and realized continuous video reconstruction. Recently, learning-based methods have made great breakthroughs in this task. Rebecq *et al*. [30, 31] proposed the first learning-based video reconstruction method, namely E2VID, which employs Recurrent Neural Networks (RNNs) to improve the quality of output frames. Scheerlinck *et al*. [33] replaced LSTMs with GRUs to get an efficient reconstruction network with much fewer parameters. Later, Weng *et al*. [41] introduced self-attention mechanisms and designed a Transformer for this task. Cadena *et al*. [5] employed spatially-adaptive denormalization (SPADE) layers and proposed SPADE-E2VID to improve the quality of reconstructed frames. SSL-E2VID [28] proposed a self-learning approach for this task to alleviate the dependence on labeled datasets. Zhu *et al*. [48] developed a novel framework based on spiking neural networks (SNNs) to achieve comparable performance with less energy consumption. More recently, Ercan *et al*. [8] introduced hypernetworks and dynamic convolutions to generate per-pixel adaptive filters that combine information from events and previously reconstructed images. Gao *et al*. [9]

proposed to leverage the bidirectional temporal information in event sequences for better predictions. Lou *et al*. [26] further proposed an efficient V2V training strategy to boost the performance of E2VID utilizing large training data. Zhu *et al*. [49] introduced diffusion models to leverage the prior from events. Despite great progress made by these methods, they lack the dense information of the scene due to the sparsity of motion-triggered events. In contrast, we combine aperture-modulation-triggered events to complement.

2.2. Reconstruction with intensity-triggered events

There are several works exploring the modulation of light intensity received by the sensor to trigger events. They can be divided into active lighting modulation and aperture modulation. Chen *et al*. [6] utilized the event streams triggered in the split second when the light is turned on for indoor lighting estimation and alleviated the problem of intensity-distance ambiguity. Han *et al*. [13] proposed the concept of transient event frequency (TEF) when the light is turned on, and achieved more accurate and stable estimations of irradiance values of the scene. Both of them need actively controlled lights, which are only practical in indoor circumstances. Bao *et al*. [1] proposed a temporal mapping photography for event cameras, where they employed Transmittance Adjustment Devices (such as aperture shutters) for brightness modulation, and employed the timestamp of the first positive event to reconstruct a dense intensity map. Later, they fused the estimated intensity map with low-light images and proposed a low-light image enhancement pipeline [36]. However, these methods only focused on static scenes and did not ensemble motion-triggered events for video reconstruction.

3. Method

In this section, we first formulate motion-triggered and aperture-modulation-triggered events in Sec. 3.1. Then we introduce our framework for the video reconstruction with aperture modulation and its key components in Sec. 3.2. Our training details are illustrated in Sec. 3.3.

3.1. Formulation

Motion-triggered events. An event camera outputs a spatio-temporal stream of polarity changes. An event signal $e = (t, \mathbf{r}, p)$ is triggered whenever the logarithmic irradiance changes exceed the preset threshold C:

$$\left| \log \frac{\mathbb{I}(\mathbf{r}, t) + I_{\text{dark}}}{\mathbb{I}(\mathbf{r}, t - \Delta t) + I_{\text{dark}}} \right| \geq C, \tag{1}$$

where $\mathbb{I}(\mathbf{r}, t)$ and $\mathbb{I}(\mathbf{r}, t - \Delta t)$ represents the pixel irradiance at pixel $\mathbf{r}$ at time t and $t - \Delta t$, respectively. The polarity $p \in \{+1, -1\}$ represents the direction of irradiance changes. I_{dark} is the dark-current term with stochastic fluctuations, which is typically much smaller than the pixel irradiance under normal illumination.

Assume that the light conditions remain unchanged, $\mathbb{I}(\mathbf{r}, t_0)$ is the irradiance at a reference timestamp t_0, and there are N_e events triggered at pixel $\mathbf{r}$ between t_0 and t. Then the irradiance at anytimestamp t can be derived as:

$$\mathbb{I}(\mathbf{r}, t) = \mathbb{I}(\mathbf{r}, t_0) \cdot \exp(\mathbf{S}(t_0, t)), \tag{2}$$

where

$$\mathbf{S}(t_0, t) = \begin{cases} C \cdot \sum_{n=1}^{N_e} p_n, & t > t_0, \\ \frac{1}{C \cdot \sum_{n=1}^{N_e} p_n}, & t < t_0. \end{cases} \tag{3}$$

Video reconstruction from pure motion-triggered events is an ill-posed problem as none of the reference irradiance $\mathbb{I}(\mathbf{r}, t_0)$ is known throughout the capture process. Because the models can only infer relative irridiance changes, *i.e*., $\mathbf{S}(t_0, t)$ from the event stream, the static regions with no events triggered are hard to reconstruct.

Aperture-modulation-triggered events. When we open the aperture from zero, the contribution of the dark current in Eq. (1) will become non-negligible [1, 15]. We assume that the transmittance rate function of our aperture shutter is $\text{TR}(t)$ whose value range is $[0, 1]$. Then the irradiance at pixel $\mathbf{r}$ can be calculated as:

$$\mathbb{I}(\mathbf{r}, t) = \mathbb{I}_{\max}(\mathbf{r}) \cdot \text{TR}(t), \tag{4}$$

where $\mathbb{I}_{\max}(\mathbf{r})$ is the irradiance when the transmittance is 1.

Based on Eq. (1), if we increase the transmittance from 0, *i.e*., $\text{TR}(0) = 0$, the first positive event (FPE) triggered at pixel $\mathbf{r}$ will occur at time $t^\star(\mathbf{r})$, which satisfies:

$$\log \frac{\mathbb{I}_{\max}(\mathbf{r})\text{TR}(t^\star(\mathbf{r})) + I_{\text{dark}}}{I_{\text{dark}}} = C. \tag{5}$$

Therefore, an inverse proportion relation between $\mathbb{I}_{\max}(\mathbf{r})$ and $\text{TR}(t^\star(\mathbf{r}))$ holds:

$$\mathbb{I}_{\max}(\mathbf{r}) = \frac{(e^C - 1) \cdot I_{\text{dark}}}{\text{TR}(t^\star(\mathbf{r}))} \propto \frac{1}{\text{TR}(t^\star(\mathbf{r}))}. \tag{6}$$

Equation (6) indicates that we can use the aperture-modulation-triggered events to calculate the FPE timestamp $t^\star(\mathbf{r})$, and then estimate the dense intensity map $\mathbb{I}_{\max}(\mathbf{r})$ for each pixel $\mathbf{r}$ in the scene. These intermediate dense intensity maps can complement the motion-triggered events, serving as the reference $\mathbb{I}(\mathbf{r}, t_0)$ in Eq. (2). Note that the above derivations are invalid for events triggered by aperture closure, as the initial pixel voltage is unknown [1, 36]. Therefore, the "aperture-modulation-triggered events" exclusively refer to events triggered by aperture opening and do not include those triggered by closing throughout this paper.

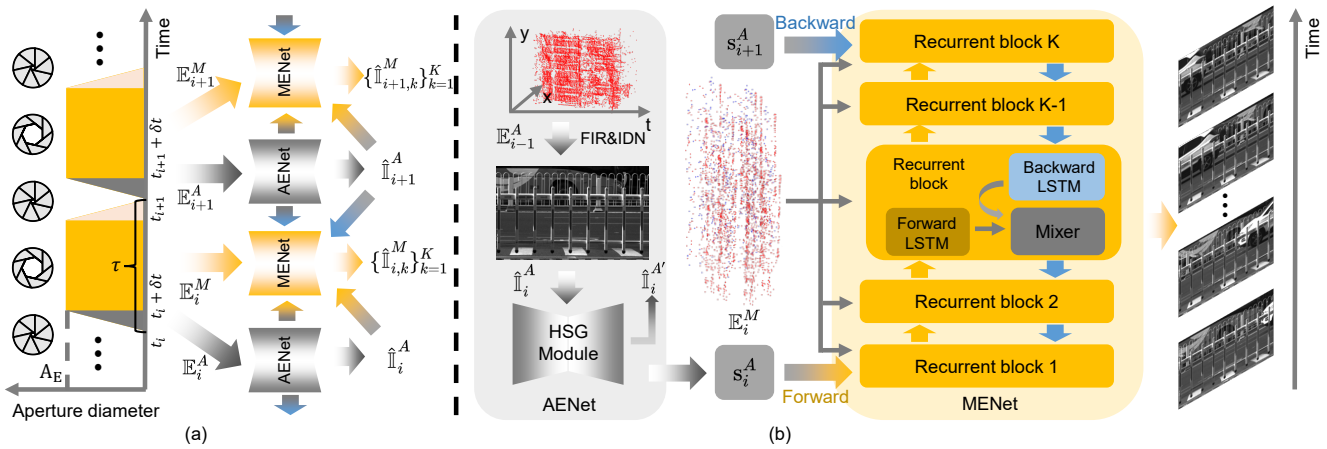

Figure 2. The overview of our AE2VID framework. (a) We periodically open the aperture to A_E with an interval τ. Note that the proportion of δt and τ does not reflect the actual values, as in practice we use a relatively small δt compared with τ (Sec. 4.1). $\mathbb{E}_i^A$ and $\mathbb{E}_i^M$ are extracted from each time window $[t_i, t_{i+1}]$, and fed into AENet and MENet, respectively. AENet predicts dense references $\hat{\mathbb{I}}_i^A$, which are also fed into MENet for the frame sequence $\{\hat{\mathbb{I}}_{i,k}^M\}_{k=1}^K$. (b) The detailed architecture of AENet and MENet. AENet contains an FIR module for initial intensity prediction, an IDN module for denoising, and an HSG module that predicts a hidden state s_i^A using features extracted from the frame. MENet is a bidirectional network consisting of the recurrent block with bidirectional LSTMs and a pixel-wise mixer. It reconstructs frames from dense references, hidden states, and event streams.

3.2. AE2VID framework

According to Eq. (2), we may only need to open the aperture once at t_0 to get a reference intensity map, and predict all the other frames with the motion-triggered events. However, we observe that as the temporal span increases, the predictions tend to exhibit larger errors at timestamps farther away from t_0. As shown in Fig. 2 (a), to alleviate the accumulation of prediction errors, we propose a periodic aperture closing and reopening mechanism that intermittently resets the observation window, thereby providing additional supervisory signals for more stable predictions.

Suppose that we open the aperture from zero at timestamps $t_i, i = 0, ..., N$, with a globally equivalent interval $\tau = t_{i+1} - t_i$. Then we get N observation windows, each corresponding to time $[t_i, t_{i+1}]$. If each opening process takes time δt, the events triggered during one temporal window could be further split into aperture-modulation-triggered events $\mathbb{E}_i^A$ for $[t_i, t_i + \delta t]$, motion-triggered events $\mathbb{E}_i^M$ for $[t_i + \delta t, t_{i+1} - \delta t]$, and aperture-closing-triggered events $\mathbb{E}_i^C$ for $[t_{i+1} - \delta t, t_{i+1}]$. Through our experiments, we find that $\mathbb{E}_i^C$ are usually noisy and provide negligible information about the scene, neither conveying dense intensity details nor motion cues. Moreover, interpolation between frames can effectively compensate for the information missing during this interval. Therefore, we simply discard these events and concentrate on the first two categories.

As illustrated in Fig. 2 (b), to facilitate more effective use of dedicated modules for handling the two types of events, our whole pipeline includes two key components: Aperture-triggered Event Network (AENet) and Motion-triggered Event Network (MENet). Based on Eq. (6), AENet takes $\mathbb{E}_i^A$ as inputs, and outputs intermediate dense intensity predictions $\hat{\mathbb{I}}_i^A$. Additionally, to provide dense reference for MENet, AENet also outputs hidden states s_i^A. According to Eq. (2), MENet takes motion-triggered events $\mathbb{E}_i^M$, intensity predictions $\hat{\mathbb{I}}_i^A$, and hidden states s_i^A from AENet, and outputs a sequence of images $\{\hat{\mathbb{I}}_{i,k}^M\}_{k=1}^K$ where K is the frame number of each sequence. They can be formulated as:

$$\begin{aligned} \hat{\mathbb{I}}_i^A, \mathrm{s}_i^A &= \mathrm{AENet}(\mathbb{E}_i^A), \\ \{\hat{\mathbb{I}}_{i,k}^M\}_{k=1}^K &= \mathrm{MENet}(\mathbb{E}_i^M, \mathrm{s}_i^A, \mathrm{s}_{i+1}^A, \hat{\mathbb{I}}_i^A, \hat{\mathbb{I}}_{i+1}^A). \end{aligned} \tag{7}$$

AENet. The goal of AENet is to predict intermediate dense intensity maps and hidden states from aperture-modulation-triggered events. In order to fully exploit the information encoded in $\mathbb{E}_i^A$, our AENet consists of three modules: (1) FPE-based Intensity Reconstruction (FIR), (2) Image Denoising (IDN), and (3) Hidden State Generation (HSG).

In the FIR module, we first build a temporal matrix from the FPE of each pixel in $\mathbb{E}_i^A$. Subsequently, we reconstruct initial intensity images $\hat{\mathbb{I}}_i^{FIR}$ based on Eq. (6). As the temporal matrix may contain much noise, we further employ a network to denoise the initial image into $\hat{\mathbb{I}}_i^A$.

We adopt SwinIR [22] as our denoising network, and employ the checkpoint pretrained by [1]. Since this checkpoint introduces an additional super-resolution effect, we further downsample the output to restore the original resolution.

To provide robust initialization and reference for MENet from $\mathbb{E}_i^A$, we further introduce the HSG module. For better feature alignments, we adopt the same structure as the for-

ward LSTM of MENet's recurrent block for HSG module. Given the denoised frame $\hat{\mathbb{I}}_i^A \in \mathbb{R}^{H\times W}$, where (H, W) is the sensor resolution, we replicate it in channel axis for b times to form a frame voxel $V_i^A \in \mathbb{R}^{b\times H\times W}$ where b in the number of bins for event voxels. The HSG module then produces the hidden state s_i^A from V_i^A. Besides, to better supervise the training of HSG, an additional pseudo-frame $\hat{\mathbb{I}}_i^{A'}$ is also predicted, which can be formulated as:

$$\hat{\mathbb{I}}_i^{A'}, \mathrm{s}_i^A = \mathrm{HSG}(V_i^A). \tag{8}$$

MENet. For reconstructing temporally consistent frames from motion-triggered events and dense references provided by AENet, we propose a recurrent network MENet. The backbone of MENet is based on E2VID [30, 31], which employs convolutional LSTM [34] as the recurrent unit. While the majority of existing video reconstruction frameworks operate in a unidirectional fashion, we observe that the absence of backward temporal information notably impairs the fidelity of static background reconstruction under long-term temporal dependencies. To mitigate this issue, we propose a bidirectional pipeline that more effectively leverages temporal correlations in the event streams as well as guidance from AENet. Note that strictly speaking, the reference frames fed into MENet should be at timestamps $t_i + \delta t$ and $t_{i+1} - \delta t$ since $\delta t > 0$. However, as δt is relatively small compared with τ (refer to Sec. 4.1), here we approximate $\hat{\mathbb{I}}_i^A$ as the intensity at $t_i + \delta t$ and $\hat{\mathbb{I}}_{i+1}^A$ as the intensity at $t_{i+1} - \delta t$.

For each time window $[t_i + \delta t, t_{i+1} - \delta t]$, we run a forward process from the initial hidden state s_i^A and a backward process from the terminal hidden state s_{i+1}^A. Specifically, we first convert event stream $\mathbb{E}_i^M$ into $K + 1$ voxel grids $\{V_{i,k}^M\}_{k=1}^{K+1}$. In the forward run, we sequentially feed voxel grids $V_{i,k}^M$ and hidden states from the previous iteration $\mathrm{s}_{i,k-1}^{M,\mathrm{fwd}}$ into the recurrent block $\mathcal{R}$, which outputs forward predictions $\hat{\mathbb{I}}_{i,k}^{M,\mathrm{fwd}}$. This can be formulated as:

$$\hat{\mathbb{I}}_{i,k}^{M,\mathrm{fwd}}, \mathrm{s}_{i,k}^{M,\mathrm{fwd}} = \mathcal{R}(V_{i,k}^M, \mathrm{s}_{i,k-1}^{M,\mathrm{fwd}}), \tag{9}$$

where $\mathrm{s}_{i,0}^{M,\mathrm{fwd}} = \mathrm{s}_i^A, k = 1, 2, ..., K$.

In the backward run, we first reverse the events with the approach similar to [37]. Then we feed the reversed voxel grids in a backward manner, which is:

$$\hat{\mathbb{I}}_{i,k}^{M,\mathrm{bwd}}, \mathrm{s}_{i,k}^{M,\mathrm{bwd}} = \mathcal{R}(\mathrm{rev}(V_{i,k+1}^M), \mathrm{s}_{i,k+1}^{M,\mathrm{bwd}}), \tag{10}$$

where $\mathrm{rev}(\cdot)$ denotes the reverse operator for each voxel, and $\mathrm{s}_{i,K+1}^{M,\mathrm{bwd}} = \mathrm{s}_{i+1}^A$.

To effectively utilize complementary information from both directions, we fuse forward and backward candidates $\hat{\mathbb{I}}_{i,k}^{M,\mathrm{fwd}}$, $\hat{\mathbb{I}}_{i,k}^{M,\mathrm{bwd}}$, and the reference frames $\hat{\mathbb{I}}_i^A$ and $\hat{\mathbb{I}}_{i+1}^A$ for each intermediate frame k with a lightweight pixel-wise mixer $\mathcal{M}$. The mixer predicts a weight $\alpha_{i,k} \in [0, 1]^{4\times H\times W}$:

$$\alpha_{i,k} = \mathrm{softmax}(\mathcal{M}(\hat{\mathbb{I}}_{i,k}^{M,\mathrm{fwd}}, \hat{\mathbb{I}}_{i,k}^{M,\mathrm{bwd}}, \hat{\mathbb{I}}_i^A, \hat{\mathbb{I}}_{i+1}^A, V_{i,k}^M)). \tag{11}$$

Finally, we obtain the prediction by mixing the four items:

$$\begin{aligned}\hat{\mathbb{I}}_{i,k}^M = \alpha_{i,k}^{(0)} \odot \hat{\mathbb{I}}_{i,k}^{M,\mathrm{fwd}} + \alpha_{i,k}^{(1)} \odot \hat{\mathbb{I}}_{i,k}^{M,\mathrm{bwd}} \\ + \alpha_{i,k}^{(2)} \odot \hat{\mathbb{I}}_i^A + \alpha_{i,k}^{(3)} \odot \hat{\mathbb{I}}_{i+1}^A,\end{aligned} \tag{12}$$

3.3. Training details

Loss functions. Because all observation windows share an identical processing pipeline, it suffices to apply the constraint to a single window. Accordingly, we omit subscripts i in loss functions. To supervise the HSG module to output feature-aligned hidden states, we first constrain the similarity between the pseudo frame $\hat{\mathbb{I}}^{A'}$ output by HSG and the reconstructed frame $\hat{\mathbb{I}}^A$ from events with an ℓ_1 loss $\|\hat{\mathbb{I}}^{A'} - \hat{\mathbb{I}}^A\|_1$. Additionally, for the quality of predicted image sequence $\hat{\mathbb{I}}_k^M$, we utilize a combination of ℓ_1 loss, VGG perceptual loss [18], and temporal consistency loss [21]. The first two loss functions are used for regularizing the fidelity between predicted frames $\hat{\mathbb{I}}_k^M$ and ground truth $\mathbb{I}_k^M$:

$$\mathcal{L}_{\mathrm{rec}}^k = \|\hat{\mathbb{I}}_k^M - \mathbb{I}_k^M\|_1 + d(\hat{\mathbb{I}}_k^M, \mathbb{I}_k^M), \tag{13}$$

where $d(\cdot)$ denotes the LPIPS distance. The temporal consistency loss is for reducing the temporal artifacts. However, as we observe that this term may introduce dirty-window artifacts [26], only the latter half frames are applied. Thus, the final loss over K frames per window is:

$$\mathcal{L} = \|\hat{\mathbb{I}}^{A'} - \hat{\mathbb{I}}^A\|_1 + \sum_{k=0}^{K} \mathcal{L}_{\mathrm{rec}}^k + \lambda_{\mathrm{TC}} \sum_{k=L_0}^{K} \mathcal{L}_{\mathrm{TC}}^k, \tag{14}$$

where we set $K = 20$, $L_0 = 10$, and $\lambda_{\mathrm{TC}} = 1$.

Training datasets. Aligning with the previous method [30, 31], we employ the **synthetic dataset** generated by the event simulator ESIM [14], which contains 1,000 2-second sequences for training our framework. However, as their datasets only contain globally homographic motion, we find that solely training on these data leads to a lack of generality and unsatisfactory performance. In order to add to the motion diversity, we further synthesize our own training dataset with Blender [3], where we randomly select objects to move in the foreground, and wrap their surfaces with images sampled from the MS-COCO dataset [23]. The generated videos are further fed into ESIM to simulate events. Our newly synthesized dataset is composed of 500 1-second sequences, resulting in more than 40 minutes of training data in total.

Implementation details. Our framework is implemented using Pytorch [29] and runs on a single NVIDIA GeForce

RTX 4090 GPU. We choose AdamW [25] as the optimizer. We augment our training data using random 2D rotations ($\pm 20°$), horizontal and vertical flips, and random cropping (128×128). The input events for MENet are transormed into voxel grids with $b = 5$ bins. We initialize the weights of both the HSG module and the recurrent block of MENet with the pretrained checkpoint of V2V-E2VID [26]. The training is conducted in two stages. In the first stage, we freeze MENet and finetune the HSG module for 10 epochs with a batch size of 4 to enforce the hidden states output by HSG to align with MENet. In the second stage, we finetune the whole framework for 10 epochs with a batch size of 4. For each training stage, we adopt a cosine learning rate decay strategy with an initial learning rate of 10^{-5} and a minimum learning rate of 10^{-7}. An additional gamma correction with parameter 1.2 is further applied for each predicted frame.

4. Experiments

In this section, we first introduce our real data capture process in Sec. 4.1. The experimental settings, including datasets, compared methods, and metrics, are elaborated in Sec. 4.2. Then we provide quantitative comparison results in Sec. 4.3, and illustrate qualitative comparisons in Sec. 4.4. Ablation studies including the advantage of bidirectional pipeline, two-stage training scheme, and the structure of AENet can be found in the supplementary material.

4.1. Aperture modulation real data capture

We capture a real Aperture-modulation-triggered and Motion-triggered Events Dataset (AMED) with the aperture modulation system shown in Fig. 1 (a). The system contains a Prophesee EVK4 event camera with resolution 1280×720 and a Computar LensConnect BH Series Variable Focal Length Lens[1] as the aperture shutter. The aperture of this lens can be modulated with a motorized module according to the predefined curves in the software.

To obtain higher-quality videos through aperture modulation in the real world, several critical aperture control parameters must be configured. Firstly, the time of the aperture opening process δt is determined jointly by the final aperture position A_E and the aperture speed v_A. Our experiments reveal that an excessively large A_E or a slow v_A prolongs the opening process, causing motion cue degradation and temporal discontinuities across frames. Conversely, when A_E is too small, some pixels fail to trigger FPE, while an overly fast v_A may produce an excessive event rate beyond the sensor's handling capacity, impairing the fidelity of predicted $\hat{\mathbb{I}}_i^A$. Secondly, the interval between two consecutive opening processes τ also plays a crucial role. A large τ may lead to prediction errors accumulating in $\hat{\mathbb{I}}_{i,k}^M$, whereas an overly small τ may destabilize the system and cause excessive information loss during aperture transitions.

After extensive experiments, we find a set of empirical parameters in our data capture process, where we set A_E as $1/4$ of the largest aperture, δt as an average of 0.13 seconds, $\tau - 2\delta t$ as 5 seconds. Besides, data with other parameter values are also collected to add diversity. Detailed discussions on these parameters can be found in supplementary material.

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

Table 1. Quantitative comparisons of event-based video reconstruction on EvAid [7]. ↑ (↓) indicates the higher (lower), the better performance. The best performances are highlighted in **bold**, and the second best in underline.

Method	MSE↓	SSIM↑	MS-SSIM↑	LPIPS↓
E2VID [30, 31]	0.227	0.471	0.345	0.639
FireNet [33]	0.160	0.493	0.350	0.608
ETNet [41]	0.051	0.602	0.457	0.486
SPADE-E2VID [5]	0.126	0.508	0.335	0.623
PAEVSNN [48]	0.123	0.511	0.301	0.624
BDE2VID [9]	0.079	0.571	0.321	0.583
V2V-E2VID [26]	0.052	0.642	0.524	**0.409**
Ours	**0.037**	**0.707**	**0.544**	0.411

4.2. Experimental settings

Evaluation datasets. Our evaluation experiments are conducted in two parts. Firstly, we construct **semi-real datasets** based on a recent event-based vision benchmark EvAid [7] with diverse motion patterns and a commonly used dataset HQF [35] for quantitative and qualitative comparisons. To keep consistency with real data, we select each 5 seconds of a sequence as a time window. Within each window, the first frame is seen as captured by aperture opening, while the last corresponds to aperture closing. As both datasets only contain motion-triggered events, we synthesize each $\hat{\mathbb{I}}_i^{FIR}$ obtained by aperture opening based on the degradation model proposed by [1]. Note that the simulation process replaces only the FIR module, and the synthesized frame will be further fed into the IDN module for denoising. The last frame is discarded following the procedure described in Sec. 3.2. After processing each sequence, we interpolate the missing frames with neighboring ones using RIFE [17]. In contrast, the compared methods directly use motion-triggered events from the dataset as their inputs. Secondly, we evaluate the performance on our **real-captured AMED dataset**. As we do not capture ground truth images, only qualitative comparisons are conducted in the paper.

Compared methods. To prove the effectiveness of our method, we compare our framework with seven state-of-the-art event-based video reconstruction methods, consisting of E2VID [30, 31], FireNet [33], ETNet [41], SPADE-E2VID [5], PAEVSNN [48], BDE2VID [9], and V2V-E2VID [26]. For all the compared methods, we use their officially released codes and pretrained checkpoints.

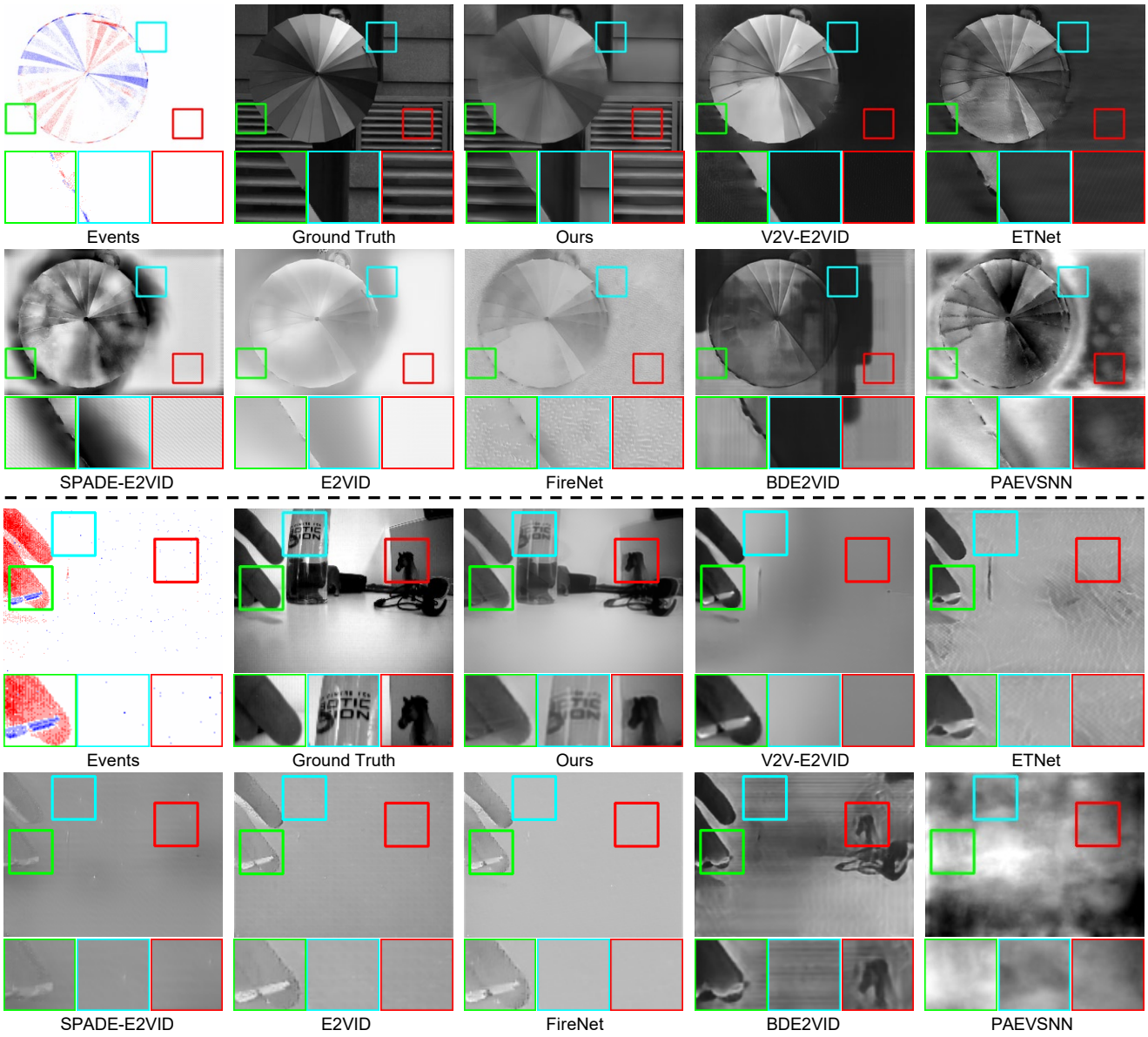

Figure 3. Qualitative experiment results on EvAid [7] and HQF [35]. The top two rows correspond to EvAid and the bottom are HQF. We compare with V2V-E2VID [26], ETNet [41], SPADE-E2VID [5], E2VID [30, 31], FireNet [33], BDE2VID [9], and PAEVSNN [48].

Metrics. To assess the quality of predictions, 4 commonly used metrics are adopted, including Mean Squared Error (MSE), Structural Similarity (SSIM) [40], Multi-Scale Structural Similarity (MS-SSIM) [39], and Perceptual Loss (LPIPS) [47]. For consistency, the calculation strategy is the same as the previous methods [26].

Table 2. Quantitative comparisons on HQF [35].

Method	MSE↓	SSIM↑	MS-SSIM↑	LPIPS↓
E2VID [30, 31]	0.187	0.475	0.325	0.511
FireNet [33]	0.096	0.520	0.387	0.456
ETNet [41]	0.043	<u>0.529</u>	0.458	<u>0.294</u>
SPADE-E2VID [5]	0.072	0.480	0.289	0.487
PAEVSNN [48]	0.098	0.496	0.371	0.484
BDE2VID [9]	<u>0.041</u>	0.523	<u>0.477</u>	**0.272**
Ours	**0.039**	**0.585**	**0.503**	0.352

4.3. Quantitative results

Quantitative results on EvAid [7] and HQF [35] datasets are shown in Table 1 and Table 2, respectively. It can be observed that our method achieves almost the best in all metrics. Specifically, we achieve more than 27.4% improvements in MSE on EvAid compared with state-of-the-art methods, demonstrating the robustness of our method across diverse

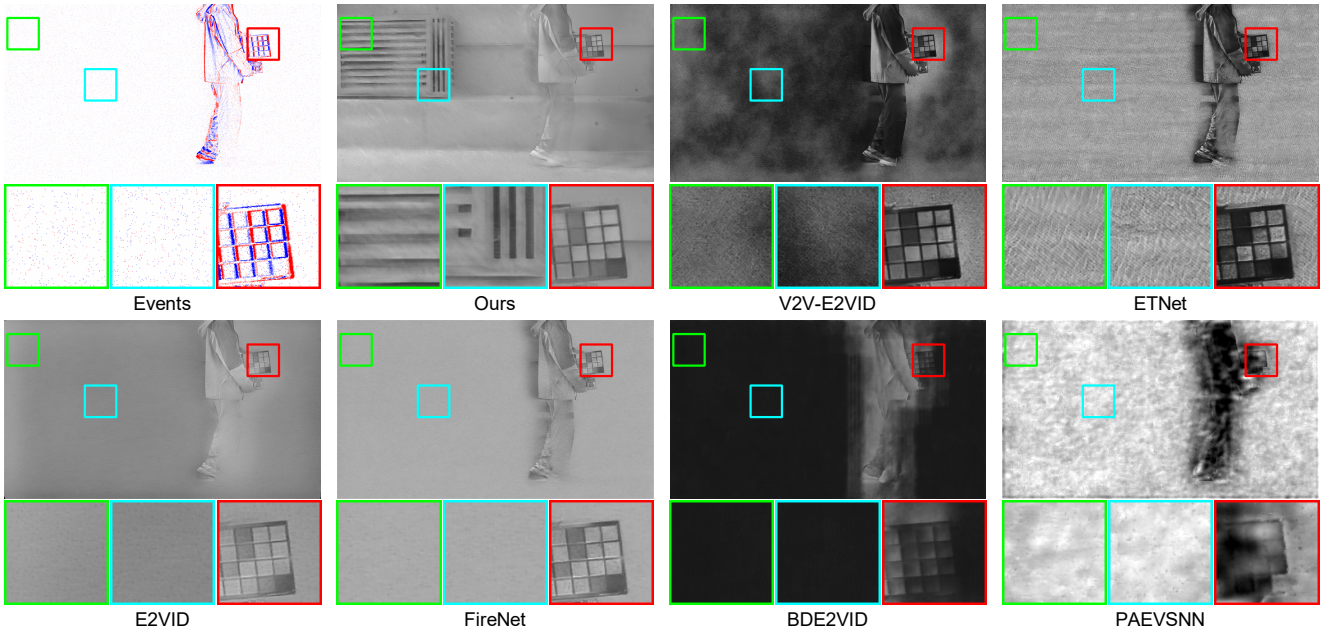

Figure 4. Qualitative experiment results on our AMED dataset with corresponding input motion-triggered events. We compare with V2V-E2VID [26], ETNet [41], E2VID [30, 31], FireNet [33], BDE2VID [9], and PAEVSNN [48].

scenarios. Although the performance increase on HQF is relatively smaller as it mainly contains global motion scenes, it also shows the effectiveness of the proposed strategy.

4.4. Qualitative results

Qualitative comparisons on EvAid [7] and HQF [35] are shown in Fig. 3. Our method consistently yields superior reconstructions of backgrounds compared with others, particularly in scenes with dominant local motion. For instance, the wall in the first group and the bottle in the second group are faithfully reconstructed by our method, whereas compared methods fail due to lack of motion-triggered events.

Furthermore, we validate the effectiveness of our method in real-world scenarios using our AMED dataset. The qualitative comparison results on AMED are presented in Fig. 4. Our method preserves the most detailed background structures while simultaneously maintaining accurate intensity information in the foreground. For example, the wall and objects in the background are best preserved by our method compared with others. Besides, we also reconstruct the details of the colorchecker with high fidelity. More qualitative results can be found in the supplementary material.

5. Conclusion

In this paper, we propose to leverage aperture-modulation-triggered events to assist event-based video reconstruction. Since motion-triggered events are usually spatially sparse, we observe that aperture modulation can provide complementary dense information about the scene. Building on this insight, we design an AE2VID framework composed of AENet and MENet for the dedicated processing of events triggered by these two sources. Furthermore, we capture a real-world AMED dataset using our modulation strategy for evaluation. Both quantitative and qualitative results demonstrate the effectiveness and superiority of our method.

Limitations. Due to our current hardware implementation limitations, we can only specify one fixed set of parameters, including A_E, v_A and τ, for each capturing process, which constrains the flexibility of the current prototype. However, when dealing with complex intensity variations and high-speed scenarios in real world, dynamically adjusting these parameters according to changes in lighting conditions and motion speeds may lead to better performance. Besides, challenging scenarios such as high-speed motion and extremely low illumination could degrade the reconstruction quality of our method. We will work on improving the adaptivity and robustness of our system in the future.

Acknowledgments

This work was supported by Beijing Natural Science Foundation (Grant No. L233024), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012), and National Natural Science Foundation of China (Grant No. 62136001, 62402014). Peiqi Duan was supported by China National Postdoctoral Program for Innovative Talents (Grant No. BX20230010) and China Postdoctoral Science Foundation (Grant No. 2023M740076).

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. 2, 3, 4, 6

[2] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[3] Blender Foundation. The Blender project - free and open 3D creation software. Accessed: 2025-09-30. 5

[4] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240 × 180 130 db 3 µs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 1

[5] 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, 6, 7, 3

[6] Zehao Chen, Qian Zheng, Peisong Niu, Huajin Tang, and Gang Pan. Indoor lighting estimation using an event camera. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[7] 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. 2, 6, 7, 8, 3, 5

[8] Burak Ercan, Onur Eker, Canberk Saglam, Aykut Erdem, and Erkut Erdem. HyperE2VID: Improving event-based video reconstruction via hypernetworks. *IEEE Transactions on Image Processing*, 33:1826–1837, 2024. 2

[9] 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. 1, 2, 6, 7, 8, 3, 9

[10] Qiyao Gao, Peiqi Duan, Hanyue Lou, Minggui Teng, Ziqi Cai, Xu Chen, and Boxin Shi. Unified reconstruction of static and dynamic scenes from events. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[11] Chengjie Ge, Xueyang Fu, Peng He, Kunyu Wang, Chengzhi Cao, and Zheng-Jun Zha. EventMamba: Enhancing spatio-temporal locality with state space models for event-based video reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 1

[12] Hanqian Han, Jianing Li, Henglu Wei, and Xiangyang Ji. Event-3DGS: Event-based 3D reconstruction using 3D gaussian splatting. In *Advances in Neural Information Processing Systems*, 2024. 2

[13] Jin Han, Yuta Asano, Boxin Shi, Yinqiang Zheng, and Imari Sato. High-fidelity event-radiance recovery via transient event frequency. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[14] Rebecq Henri, Gehrig Daniel, and Scaramuzza Davide. ESIM: an open event camera simulator. In *Proc. of Conference on Robot Learning*, 2018. 5

[15] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. V2E: From video frames to realistic DVS events. In *Proc. of Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021. 3

[16] Tiejun Huang, Yajing Zheng, Zhaofei Yu, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, et al. 1000× faster camera and machine vision with ordinary devices. *Engineering*, 25:110–119, 2023. 1

[17] 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. 6, 3

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. of European Conference on Computer Vision (ECCV)*, 2016. 5

[19] Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew J Davison. Simultaneous mosaicing and tracking with an event camera. *J. Solid State Circ*, 43:566–576, 2008. 2

[20] 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*, 2023. 2

[21] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 5

[22] Jingyun Liang, Jiezhang 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. 4, 3

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. of European Conference on Computer Vision (ECCV)*, 2014. 5

[24] Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. EventGPT: Event stream understanding with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 2

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. of International Conference on Learning Representations (ICLR)*, 2019. 6

[26] 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. 1, 3, 5, 6, 7, 8, 9

[27] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126(12):1381–1393, 2018. 2

[28] Federico Paredes-Vallés and Guido C. H. E. de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proc.*

of Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 2

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[30] 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. 1, 2, 5, 6, 7, 8, 3, 9

[31] 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. 1, 2, 5, 6, 7, 8, 3, 9

[32] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Proc. of Asian Conference on Computer Vision (ACCV)*, 2018. 2

[33] 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. 1, 2, 6, 7, 8, 3, 9

[34] 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 nowcasting. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 5, 3

[35] Timo Stoffregen, 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. 6, 7, 8, 5

[36] Lei Sun, Yuhan Bao, Jiajun Zhai, Jingyun Liang, Yulun Zhang, Kaiwei Wang, Danda Pani Paudel, and Luc Van Gool. Low-light image enhancement using event-based illumination estimation. In *Proc. of International Conference on Computer Vision (ICCV)*, 2025. 2, 3

[37] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time Lens: Event-based video frame interpolation. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[38] 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 *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[39] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, pages 1398–1402. Ieee, 2003. 7

[40] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 7

[41] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proc. of International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 6, 7, 8, 3, 9

[42] Wenhao Xu, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. CEIA: CLIP-based event-image alignment for open-world event-based understanding. In *Proc. of European Conference on Computer Vision Workshops (ECCVW)*, 2024. 2

[43] Siqi Yang, Jinxiu Liang, Zhaojun Huang, Yeliduosi Xiaokaiti, Yakun Chang, Zhaofei Yu, and Boxin Shi. Spikediff: Zero-shot high-quality video reconstruction from chromatic spike camera and sub-millisecond spike streams. In *Proc. of International Conference on Computer Vision (ICCV)*, 2025. 1

[44] Yixin Yang, Jiawei Zhang, Yang Zhang, Yunxuan Wei, Dongqing Zou, Jimmy S Ren, and Boxin Shi. Event-guided hdr reconstruction with diffusion priors. In *Proc. of International Conference on Computer Vision (ICCV)*, 2025. 2

[45] Jiajie Yu, Xing Lu, Lijun Guo, Chong Wang, Guoqi Li, and Jiangbo Qian. Event-based video reconstruction via spatial–temporal heterogeneous spiking neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):8478–8494, 2025. 2

[46] Lei Yu, Xiang Zhang, Wei Liao, Wen Yang, and Gui-Song Xia. Learning to see through with events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8660–8678, 2022. 1

[47] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7

[48] 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, 6, 7, 8, 3, 9

[49] Lin Zhu, Yunlong Zheng, Yijun Zhang, Xiao Wang, Lizhi Wang, and Hua Huang. Temporal residual guided diffusion framework for event-driven video reconstruction. In *Proc. of European Conference on Computer Vision (ECCV)*, 2024. 2, 3

AE2VID: Event-based Video Reconstruction via Aperture Modulation

Supplementary Material

Chenxu Bai[1,2*] Boyu Li[1,2*] Peiqi Duan[1,2†] Xinyu Zhou[3] Hanyue Lou[1,2] Boxin Shi[1,2,4†]

[1] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3] State Key Laboratory of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University
[4] PKU-AI[2] Robotics Joint Lab of Embodied AI

chenxu.bai@stu.pku.edu.cn {liboyu, duanqi0001, zhouxiny, 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 [7] 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 τ. 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 δt, 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 δt

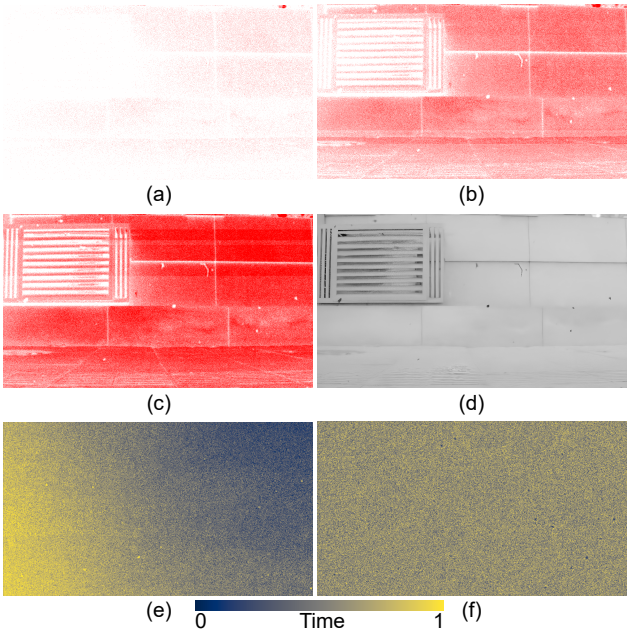

Figure 5. (a)-(c) Visualization of aperture-modulation-triggered events in chronological order. (d) Reconstructed frame from aperture-modulation-triggered events. (e) FPE temporal matrix of aperture shutter. (f) FPE temporal matrix of manual rotation.

*Equal contribution.
†Corresponding authors.

without much loss of motion cues. As δt is relatively much smaller than τ, 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 [22] 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 [34] 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 δt. However, since we have already reconstructed the frames immediately preceding the closure and those with subsequent aperture opening, we employ the RIFE model [17] 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 [30, 31], we use the `E2VID_lightweight` checkpoint; for FireNet [33], we use the `firenet_1000` checkpoint; for SPADE-E2VID [5], we use the `SPADE_E2VID` checkpoint; for V2V-E2VID [26], 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 [7] 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 ℓ_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 (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.

Table 3. Ablation study results on EvAid [7].

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	**0.707**	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	**0.037**	**0.707**	**0.544**	**0.411**

Table 4. Computation efficiency comparison results.

	Params	MACs	Time
E2VID [30, 31]	10.71 M	29.79 G	2.38 ms
FireNet [33]	37.78 K	2.46 G	1.52 ms
ETNet [41]	16.65 M	35.84 G	2.35 ms
SPADE-E2VID [5]	11.46 M	103.29 G	5.71 ms
PAEVSNN [48]	4.53 M	132.54 G	13.64 ms
BDE2VID [9]	19.35 M	54.45 G	7.16 ms
V2V-E2VID [26]	10.71 M	29.79 G	2.38 ms
Ours (w/o IDN)	53.08 M	73.53 G	4.68 ms

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 256×256, 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 τ.

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 1280×720 and a Computar LensConnect BH Series Variable Focal Length Lens[1].

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.

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

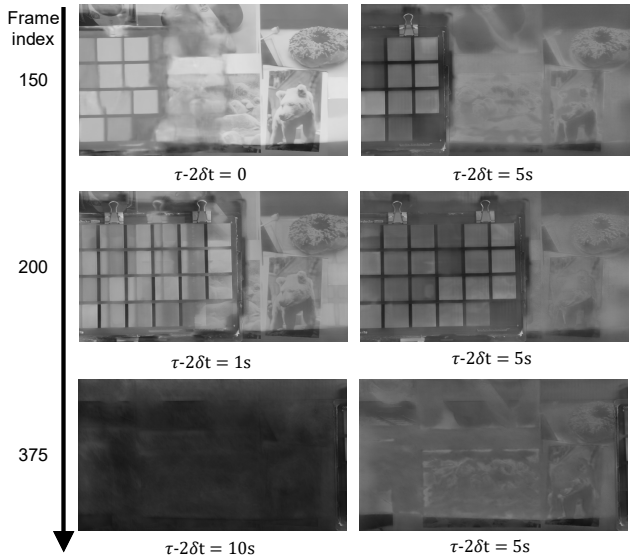

Figure 8. Reconstructed frame comparisons with different intervals. Note that each row corresponds to one different timestamp (frame index). We compare the results of $\tau - 2\delta t = 5$ seconds with $\tau - 2\delta t = 0, 1, 10$ seconds.

Interval. The length of each observation window τ 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 $\tau - 2\delta t = 1$s exhibits persistent motion artifacts, as evidenced by the colorchecker (see row 2, left). Conversely, an excessively prolonged interval of $\tau - 2\delta t = 10$s results in the degradation of background details during reconstruction (see row 3, left). Consequently, we empirically adopt $\tau - 2\delta t = 5$s to achieve a trade-off between motion fidelity and background preservation.

11. More qualitative results

More qualitative results on EvAid [7] are shown in Fig. 9, and more results on HQF [35] 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.

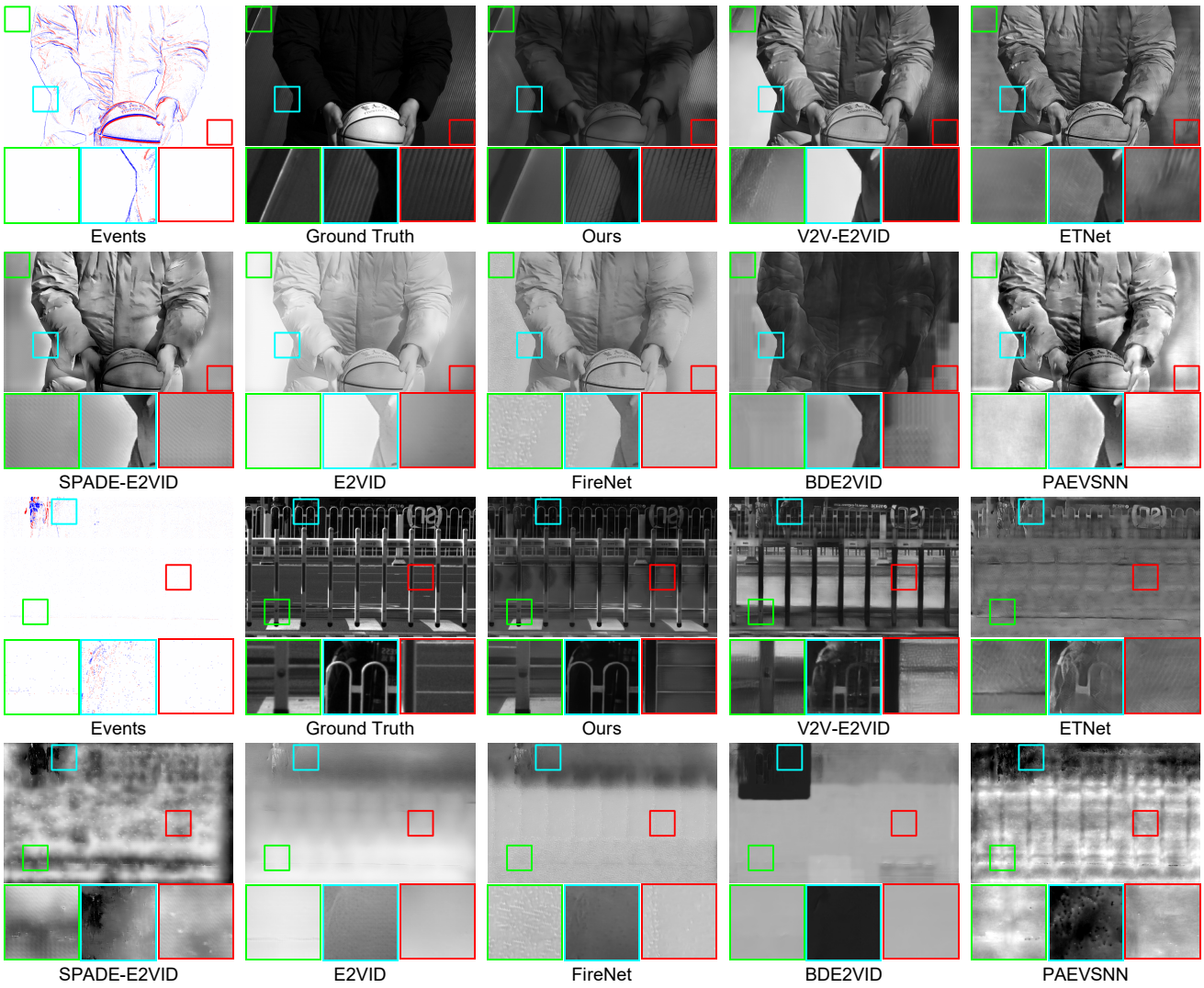

Figure 9. More qualitative experiment results on EvAid [7]. We compare with V2V-E2VID [26], ETNet [41], SPADE-E2VID [5], E2VID [30, 31], FireNet [33], BDE2VID [9], and PAEVSNN [48].

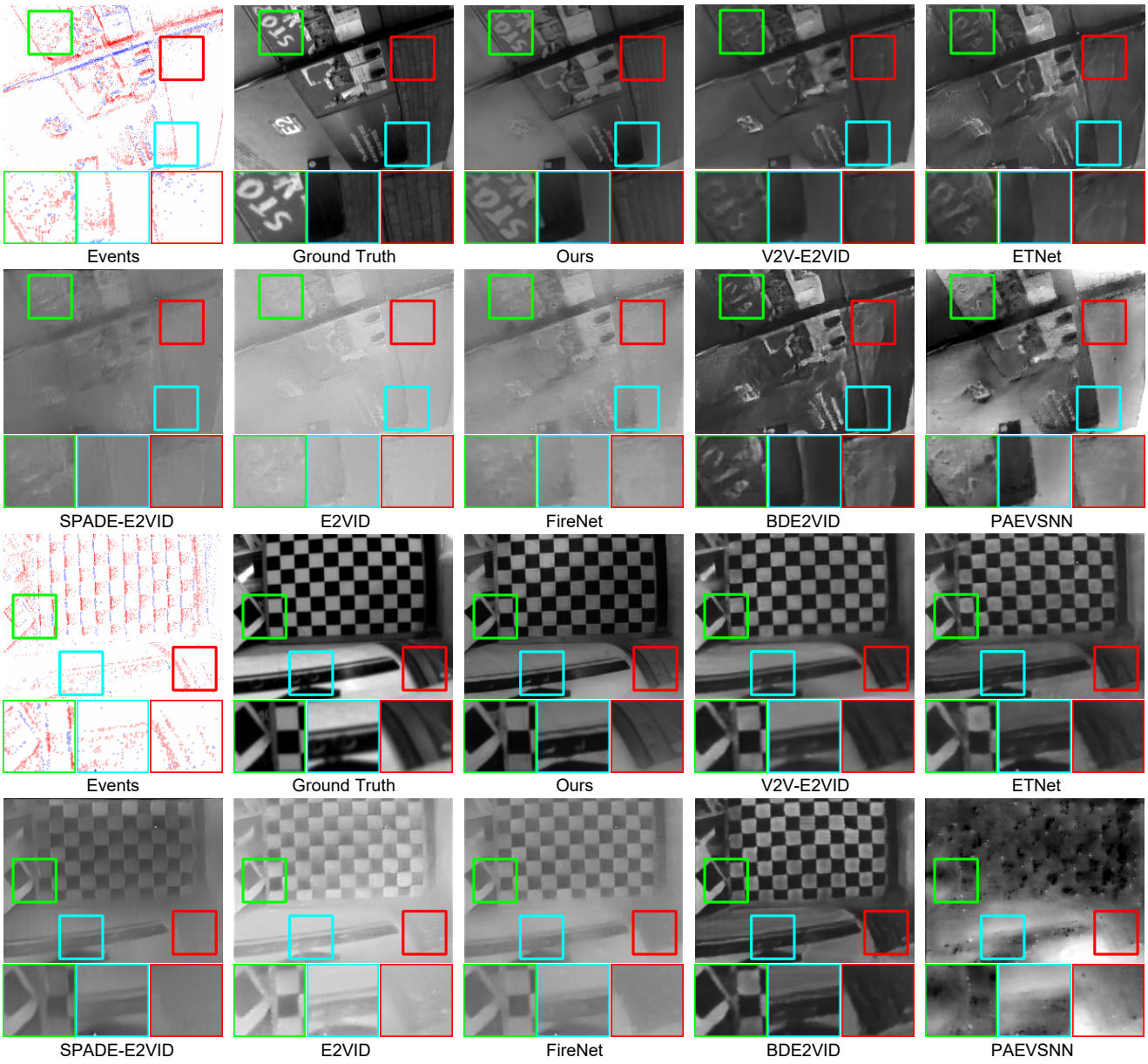

Figure 10. More qualitative experiment results on HQF [35]. We compare with V2V-E2VID [26], ETNet [41], SPADE-E2VID [5], E2VID [30, 31], FireNet [33], BDE2VID [9], and PAEVSNN [48].

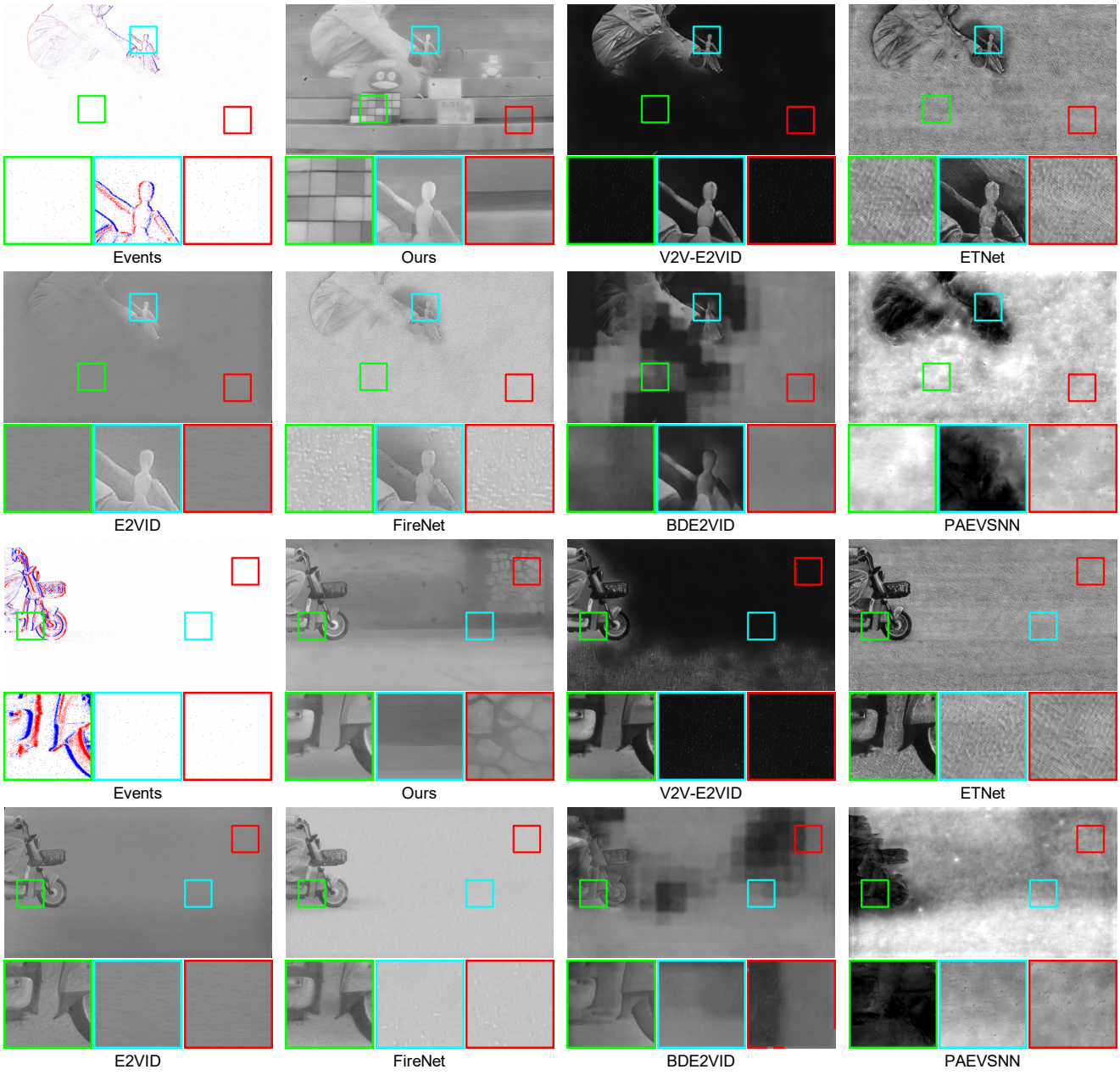

Figure 11. More qualitative experiment results on our real-captured AMED dataset (Part 1) with corresponding input motion-triggered events. We compare with V2V-E2VID [26], ETNet [41], E2VID [30, 31], FireNet [33], BDE2VID [9], and PAEVSNN [48].

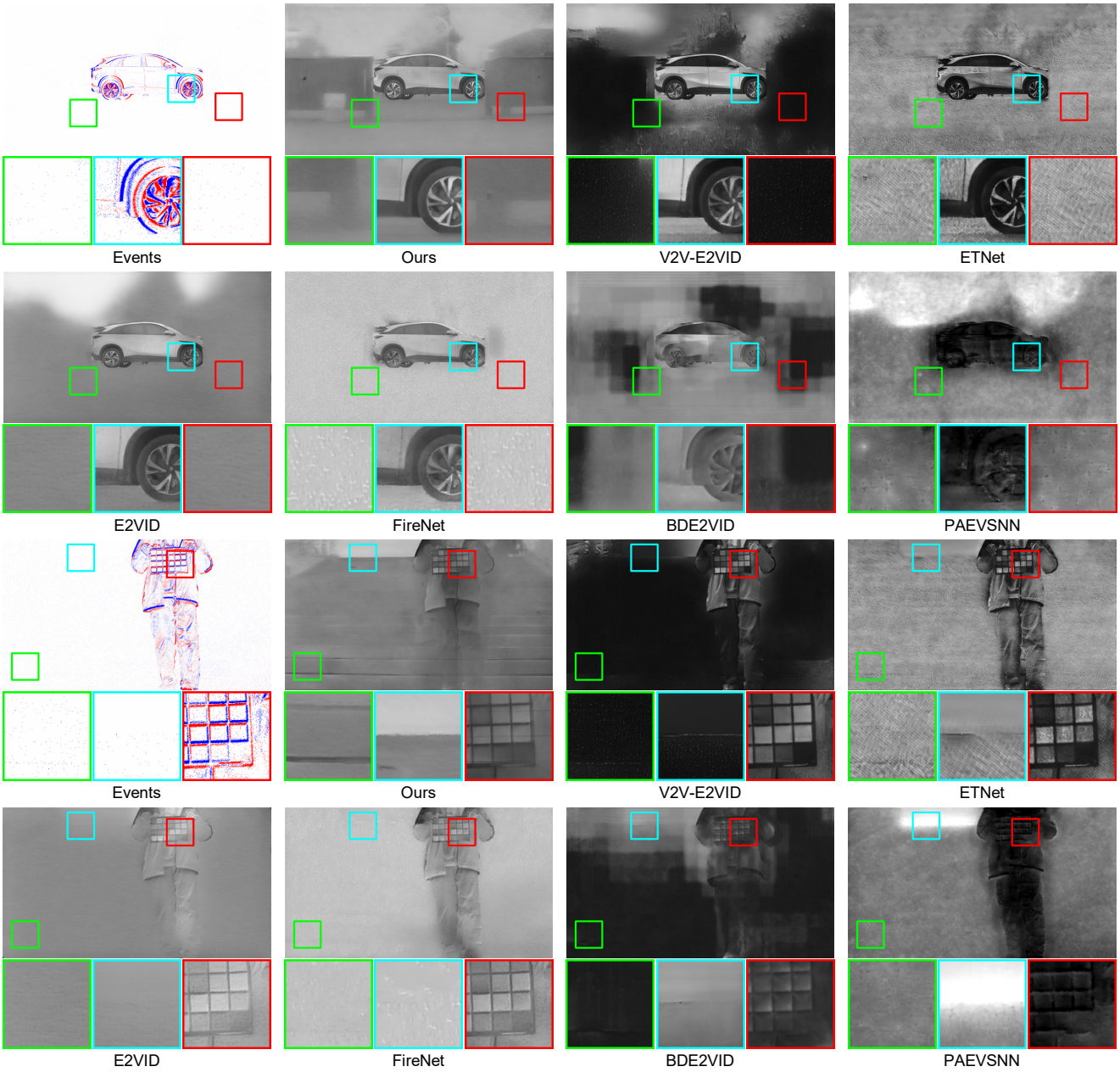

Figure 12. More qualitative experiment results on our real-captured AMED dataset (Part 2) with corresponding input motion-triggered events. We compare with V2V-E2VID [26], ETNet [41], E2VID [30, 31], FireNet [33], BDE2VID [9], and PAEVSNN [48].