

EvDiff: Event-Based Video Reconstruction using One-Step Diffusion Models

Weilun Li^{1,2,*}, Lei Sun^{2,*†}, Ruixi Gao¹, Qi Jiang¹, Yuqin Ma¹, Kaiwei Wang^{1†},
Ming-Hsuan Yang^{3,4}, Luc Van Gool², and Danda Pani Paudel²

¹ Zhejiang University, Hangzhou, China

² INSAIT, Sofia University “St. Kliment Ohridski”, Sofia, Bulgaria

³ UC Merced, CA, USA

⁴ Google DeepMind, London, United Kingdom

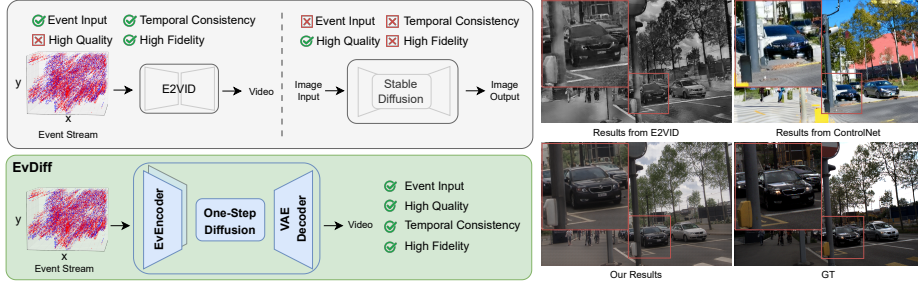


Fig. 1: Our EvDiff can reconstruct real high-quality video streams from monochrome event streams, while maintaining both fidelity and realism. Compared with Ground-Truth (GT), our result shows a higher dynamic range.

Abstract. As neuromorphic sensors, event cameras asynchronously record changes in brightness as streams of sparse events with the advantages of high temporal resolution and high dynamic range. Reconstructing intensity images from events is a highly ill-posed task due to the inherent ambiguity of absolute brightness. Early methods generally follow an end-to-end regression paradigm, directly mapping events to intensity frames in a deterministic manner. While effective to some extent, these approaches often yield perceptually inferior results and struggle to scale up in model capacity and training data. In this work, we propose EvDiff, an event-based diffusion model that follows a surrogate training framework to produce high-quality videos. To reduce the high computational cost of high-frame-rate video generation, we design an event-based diffusion model that performs only a single forward diffusion step, equipped with a temporally consistent EvEncoder. Furthermore, our novel Surrogate Training Framework eliminates the dependence on paired event-image datasets, allowing the model to leverage large-scale image datasets for higher capacity. The proposed EvDiff is capable of generating high-quality colorful videos solely from monochromatic event streams. Experiments on

* Equal contribution. Work mainly done during Weilun Li’s internship at INSAIT.

† Corresponding authors: leo_sun@zju.edu.cn, wangkaiwei@zju.edu.cn

real-world datasets demonstrate that our method strikes a sweet spot between fidelity and realism, outperforming existing approaches on both pixel-level and perceptual metrics. The code will be released publicly.

1 Introduction

Inspired by the human visual system, the Silicon Retina [39] introduced the foundation of perceptual sensing with neuromorphic cameras, commonly referred to as Dynamic Vision Sensors (DVS) or event cameras. Unlike conventional frame-based sensors, event cameras asynchronously record local changes in light intensity as streams of discrete events. Owing to the sparse nature, they provide unique advantages over their frame-based counterparts, including low power consumption, low latency, microsecond-level temporal resolution, and an ultra-high dynamic range (HDR) of up to 140 dB [15, 45].

However, due to their sparse and asynchronous nature, raw event streams cannot be directly processed by modern computer vision architectures unless specialized models such as spiking neural networks (SNNs) [70] are employed. To mitigate this gap, researchers often convert event streams into middle representations, such as voxel grids [90] or time surfaces [40]. A more direct and intuitive approach is to reconstruct video frames from events, which not only enables the visualization of high-speed and HDR videos but also bridges event cameras to the vast ecosystem of frame-based computer vision methods [49, 50].

While early works [3, 41, 53] incorporated handcrafted smoothness priors into video reconstruction, a major breakthrough was achieved by E2VID [49], which introduced a data-driven learning approach to event-to-video conversion. Subsequent studies have refined this paradigm by simplifying the network architecture [54], adding self-supervised constraints [44], and explicitly modeling the sparsity of events [6], among others. Nevertheless, most existing approaches remain constrained by the paradigm introduced by E2VID, which relies on relatively small event-image paired datasets and compact network architectures. Owing to the scarcity of large-scale training data and the inherently ill-posed nature of the reconstruction problem, these methods typically yield low-quality grayscale outputs that fall short of perceptual satisfaction for human observers.

Meanwhile, recent advances in generative models have been largely driven by the success of diffusion models [19, 51, 81]. One of the key factors behind their effectiveness lies in their scaling ability, which allows them to fully exploit large model capacities and massive high-quality datasets [55, 88].

Motivated by this trend, bringing the power of large diffusion models to event-based vision naturally becomes an appealing goal. However, two major obstacles stand in the way: ❶ *Scaling up the model size* to diffusion-based video generation frameworks is *challenging* due to the prohibitively high computational costs, as event data are inherently high-speed and the corresponding videos must be rendered at high frame rates to preserve this temporal fidelity. ❷ Compared to contemporary large-scale image datasets, *event-image paired datasets are orders of magnitude smaller*, and their large-scale collection is nearly impossible due to

the fact that event cameras are not yet widely adopted. Simulators [22, 48] also require high-frame-rate videos as inputs, which are limited in availability.

In this work, aiming to bring the powerful large foundational model to event-to-video reconstruction, we introduce EvDiff. At the model level, we adopt the Stable Diffusion 3 (SD3) model [14] as our base model. To address the problem of overwhelming computation burden to produce high-framerate video from events, we propose EvEncoder to encode events and temporal information into latent presentations, following a one-step diffusion model with SD3 as base model. At the data level, to leverage the large-scale data to feed the model, we design a Surrogate Training Pipeline to use the low-quality coarse reconstruction results to bridge the event data and image data. To be more specific, first, we design an E2VID-style Degradation Model to synthesize E2VID results from high-quality images, and we train our one-step EvDiff with the synthesized E2VID results (coarse results) and their high-quality counterparts in Place 365 [88], a large-scale image dataset with 1.8 million images. Next, we distill the E2VID and VAE encoder to our EvEncoder and then we finetune the whole model. By utilizing the Surrogate Training Pipeline, we extend the training data from the event-image-paired dataset to a much broader scale. Experiments on real-world datasets are conducted, and our EvDiff produces higher-quality chromatic videos than regression-based methods. As shown in Fig. 1, compared with ControlNet-based counterparts [86], our method is more efficient, exhibits stronger temporal consistency, and produces higher-fidelity results without relying on prompts derived from ground-truth images. Moreover, even when compared directly with the ground-truth images, our reconstructed videos often display a higher dynamic range, owing to the inherent sensing characteristics of event cameras.

The main contributions of this work are:

- We revisit the event-based video reconstruction, and propose to transfer the capabilities of large-scale diffusion models to event-based video reconstruction, achieving high reconstruction fidelity and realism while maintaining efficient inference.
- We design a novel Surrogate Training Pipeline along with the E2VID-style Degradation Model that serves as a bridge between scarce event-video paired data and large-scale natural image datasets, enabling effective model training at scale.
- Experiments show that our method reconstructs faithful and chromatic videos from monochromatic events only, competing favorably against state-of-the-art methods and the ControlNet-based counterpart.

2 Related Work

2.1 Event-Based Video Reconstruction

Due to their unique advantages, such as high temporal resolution, low latency, and robustness under challenging illumination, event cameras have found applications across a wide range of vision tasks, including SLAM, feature tracking,

image deblurring [24, 26–29, 29, 63, 66–68, 73, 80], low-light enhancement [30, 64], video interpolation [7, 18, 25, 33, 43, 65, 69, 71, 72, 87], and other tasks [2, 8, 10, 37, 82]. In addition to these tasks, we investigate the use of event cameras for video reconstruction [3, 6, 17, 41, 44, 49, 50, 53, 54, 75], *i.e.*, recovering intensity frames from event streams, in this work.

Early approaches [3, 41, 53] relied on carefully designed physical models with integration or filtering techniques, often requiring additional prior information. With the advent of deep learning, reconstruction quality improved significantly. A milestone was E2VID [49, 50], which combined a U-Net with temporal modules and achieved strong results at moderate computational cost. Later works such as FireNet [54] explored lightweight architectures, while E2VID+ and FireNet+ [61] addressed the shortage of training data by introducing synthetic events generated with ESIM [48], a strategy widely adopted in subsequent research. Following E2VID-style variants (*e.g.*, SPADE-E2VID [5], Hyper-E2VID [13], SCSE-E2VID [38], Sparse-E2VID [6]) introduced different architectural modules, typically yielding incremental improvements. Other works have explored event-based methods in resource-constrained scenarios, either using spiking neural networks or self-supervised learning [44] paradigms. The aforementioned works follow the same end-to-end training paradigm as E2VID, which typically results in relatively low-quality grayscale reconstructions with visually noticeable artifacts. Recent endeavors have sought to move beyond this paradigm. For example, diffusion-based E2VIDiff [31] proposes to condition Stable Diffusion [51] on events for the first time, producing chromatic reconstructions but with limited fidelity, leading to substantial deviations from the ground truth (GT).

2.2 One-Step Diffusion Models

Diffusion models [19, 46, 57, 59, 60, 62], which generate data by iteratively denoising random noise through a learned stochastic process, have recently achieved remarkable success as state-of-the-art deep generative models [81]. When scaled up with billions of parameters and trained on massive natural image datasets [55, 76], they demonstrate impressive realism and strong generative priors, enabling high-quality image and video synthesis [4, 47, 51, 52, 74]. However, these benefits come at the cost of heavy computational requirements, as classical diffusion processes involve hundreds of iterative denoising steps [19, 58], which limit their practicality in time-sensitive and high frame-rate video applications.

To address this inefficiency, recent works have proposed one-step diffusion approaches that distill the multi-step denoising trajectory into a single forward pass [32, 83, 84]. Originally introduced in the context of image generation, these models preserve much of the fidelity and realism of full diffusion while drastically reducing inference latency. More importantly, one-step diffusion has proven particularly effective in restoration-oriented tasks such as image super-resolution [11, 78], image deblurring [35], *etc.*, where the goal is to transform degraded inputs into high-quality outputs. Compared with multi-step diffusion, one-step diffusion offers inference efficiency that aligns well with the demands

of event-to-video reconstruction, which necessitates generating high-frame-rate videos with both accuracy and speed.

3 EvDiff

Before detailing our EvDiff for event-based video reconstruction (§ 3.2), we first formulate the problem and discuss the motivation that drives our design (§ 3.1).

3.1 Motivation

Instead of capturing full frames at a fixed frame rate, event cameras only asynchronously record per-pixel brightness changes. For each pixel, an event $e = (x, y, t, p)$ is triggered when the intensity ($\mathcal{I}$) variation in log domain exceeds a preset contrast threshold C :

$$|\log \mathcal{I}(x, y, t) - \log \mathcal{I}(x, y, t - \delta t)| \geq C, \quad (1)$$

where x , y , t , and $p \in \{+1, -1\}$ denote the coordinate, timestamp, and polarity, respectively. This unique mechanism yields high temporal resolution and dynamic range but produces sparse, non-intensity data, *i.e.*, “events”.

Following the seminal event-based video reconstruction work E2VID [49, 50], existing SOTA methods typically adopt an end-to-end training paradigm, where a single model $f(\cdot)$ directly maps a sequence of events to a sequence of images with a UNet-like model. Although this pipeline delivers substantial advances over earlier optimization-based methods [41, 53], it now encounters a critical methodological bottleneck, as outlined in the aforementioned ❶ and ❷.

As a result, E2VID-style models exhibit limited ability to recover fine spatial structures and realistic textures, leading to blur, ghosting, and artifacts.

By looking deeper into the “degradation pattern” of these results, we observe that E2VID-style models tend to produce relatively stable artifacts: characteristic combinations of blur, edge disruption, and block-like distortions. This stems from the inherent initial condition ambiguities. Therefore, if we can construct a degradation space that simulates these artifacts on images and leverage a powerful diffusion prior for reconstruction, it would bridge the event-to-video conversion task with large-scale datasets and enable large-parameter model training.

Motivated by this, we reformulate the task as

$$\{\mathbf{I}'_i\}_{i=1}^N = f_1(\{e_{\Delta t}\}; \Theta_1), \quad \{\hat{\mathbf{I}}_i\}_{i=1}^N = f_2(\{\mathbf{I}'_i\}; \Theta_2), \quad (2)$$

where $\{\mathbf{I}'_i\}_{i=1}^N$ are the degraded frames reconstructed from events, and $f_2(\cdot; \Theta_2)$ is a high-capacity generative model pretrained on large-scale image data. This decomposition provides two key advantages: it 1) removes the reliance on scarce paired event–image datasets, and 2) enables the use of large diffusion models with visual priors to achieve high-quality and perceptually consistent reconstructions.

This reformulation establishes the foundation of our framework and training strategy, described in § 3.2 and § 3.3, respectively.

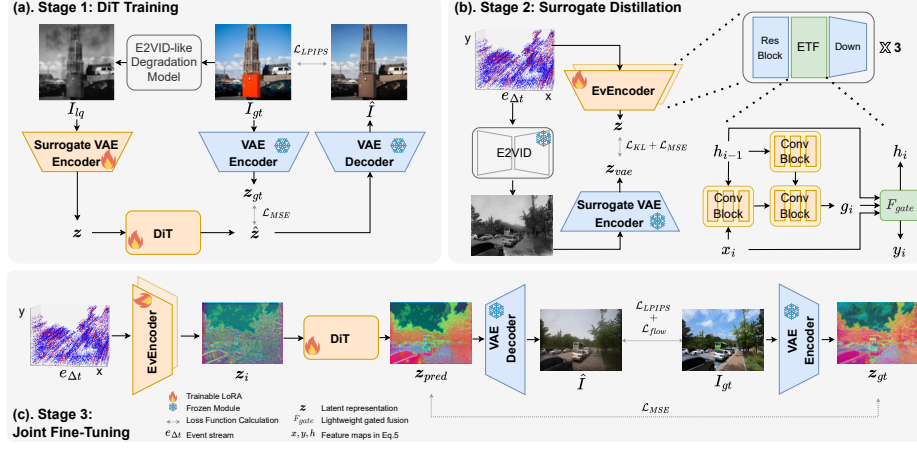


Fig. 2: The proposed Surrogate Training Pipeline. Stage 1: We train a DiT model and Surrogate VAE Encoder with LQ-HQ pairs; Stage 2: The Surrogate VAE Encoder is distilled into EvEncoder; Stage 3: We finetune the whole EvDiff model.

3.2 General Architecture of EvDiff

Following the generic form from Eq. 2, we propose EvDiff to bring the powerful diffusion prior for high-quality event-based video reconstruction. EvDiff consists mainly of two parts: a EvEncoder $\mathcal{E}_{event}$ and a one-step diffusion model OSDiff:

$$\{z_i\}_{i=1}^N = \mathcal{E}_{event}(\{e_{\Delta t}\}), \quad (3)$$

$$\{\mathbf{I}_i\}_{i=1}^N = \mathcal{D}(\text{OSDiff}(\{z_{vid}\})). \quad (4)$$

where $\mathcal{D}$ is the VAE decoder, and $\{z_i\}_{i=1}^N$ represents the latent representation. The training strategy will be detailed in §3.3. During inference, however, EvDiff operates in a single-stage, end-to-end manner, as shown in Fig. 1.

EvEncoder. Given an input event stream $\{e_{\Delta t}\}$, where $\{e_{\Delta t}\} = \{e_i \mid t_i \in [t, t + \Delta t]\}$ denotes all events occurring within the temporal window Δt , we first encode them into latent representation. To be specific, asynchronous event streams are first converted to synchronous voxel grids [90] $\mathbf{V} \in \mathbb{R}^{H \times W \times T}$ which contain temporal information in T channel. The resulting voxels are encoded by the EvEncoder to obtain latent representation for further processing. The most intuitive idea to reconstruct the videos with T frames with a diffusion model is to send all events into the diffusion model with ControlNet [86] for self-attention, similar to SVD [4], with a computational complexity of $\mathcal{O}((THW)^2 \cdot C)$. However, one of the most significant advantages of event cameras is the high temporal resolution. Thus, event-based video reconstruction methods typically focus on generating high-speed and high-frame-rate videos that conventional cameras cannot capture [50], which substantially increases the temporal length T and consequently leads to unacceptable computational overhead.

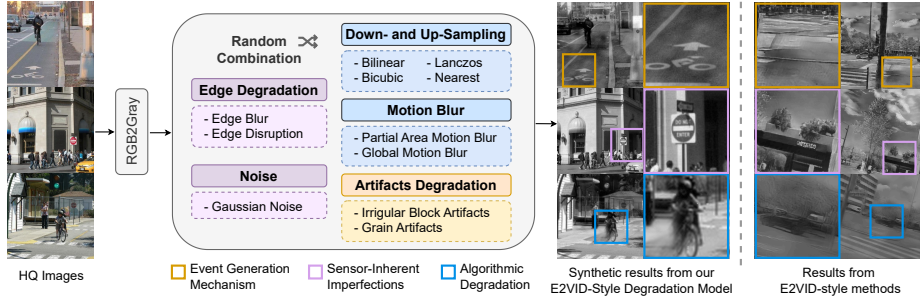


Fig. 3: Overview of the proposed E2VID-Style Degradation Model. HQ images sampled from the Places365 [88] dataset are cropped to 512×512 and then processed by our degradation model to generate corresponding LQ images. For comparison, E2VID-style results from the BS-ERGB dataset are shown on the right.

Therefore, we design a recurrent architecture [20, 56] in EvEncoder to efficiently keep the temporal information flow. For efficiency, we propose the Efficient Temporal Fusion (ETF) module in EvEncoder, which employs a lightweight gated fusion mechanism [56], as shown in Fig. 2 (b). Each ETF module contains three convolutional blocks to extract features from both historical and current data, generating a gate weight that adaptively balances their contributions:

$$\mathbf{y}_i = \mathbf{g}_i \odot \mathbf{x}_i + (1 - \mathbf{g}_i) \odot \mathbf{h}_{i-1}, \quad (5)$$

where $\mathbf{x}_i$ and $\mathbf{h}_{i-1}$ denote the current and previous feature maps, and $\mathbf{g}_i$ represents the learned fusion gate controlling temporal blending. With the ETF module, we reduce the recomputation from $\mathcal{O}((THW)^2 \cdot C)$ to $\mathcal{O}((HW)^2 \cdot T \cdot C)$ and strike a balance between temporal information flow and computational cost.

One-Step Diffusion Model. Existing SD-based methods usually take random Gaussian noise as the starting point and require multiple diffusion steps. Like aforementioned reasons, this expensive computational cost is not compatible with event cameras. Recent works [11, 78] prove that a low-quality image can be taken as a starting point for one-step diffusion. Inspired by this, we utilize Stable Diffusion 3 (SD3) [14] with powerful natural image priors as foundation model. By aligning $\mathbf{z}_i$ with latent representation from E2VID-style degraded images (see §3.3), we conduct one-step diffusion for reconstruction:

$$\hat{\mathbf{z}}_i = \text{OSDiff}(\mathbf{z}_i) \triangleq \frac{\mathbf{z}_i - \beta_{t^*} \boldsymbol{\epsilon}(\mathbf{z}_i; t^*)}{\alpha_{t^*}}, \quad (6)$$

where $\boldsymbol{\epsilon}$ denotes a denoising network, α_{t^*} and β_{t^*} are scalar coefficients determined by the fixed diffusion timestep t^* . We simplify $\mathbf{z}_i$ as $\mathbf{z}$ as we discuss only single frame case in the rest part. Then, we have the final prediction by:

$$\hat{\mathbf{I}} = \mathcal{D}(\hat{\mathbf{z}}). \quad (7)$$

So far, we have split the typical E2VID-style paradigm into two parts and introduced large diffusion models into the event-based video reconstruction task,

Table 1: Quantitative comparison on BS-ERGB [71] and DSEC [16], with the best results highlighted in **red** and the second best in **blue**.

Method	MSE↓	SSIM↑	LPIPS↓	FID↓	FVD↓	MSE↓	SSIM↑	LPIPS↓	FID↓	FVD↓
BS-ERGB						DSEC				
E2VID _[CVPR2019]	0.1175	0.3313	0.5518	276	1688	0.1042	0.3393	0.5427	239	1541
FireNet _[WACV2020]	0.0857	0.3302	0.5328	271	1909	0.0927	0.3478	0.5670	262	1492
E2VID _[ECCV2020]	0.0717	0.3710	0.4397	238	1652	0.0881	0.2812	0.5138	248	1397
FireNet _[ECCV2020]	0.0827	0.3080	0.4903	313	1562	0.0913	0.2268	0.5655	291	1592
SPADE-E2VID _[TIP2021]	0.0891	0.3235	0.6715	347	2343	0.0601	0.4578	0.4911	234	1277
SSL-E2VID _[CVPR2021]	0.0798	0.3448	0.6169	312	1907	0.1055	0.3303	0.5565	326	1521
ET-Net _[ICCV2021]	0.0683	0.3557	0.4566	270	1569	0.0737	0.2953	0.5232	259	1496
HyperE2VID _[TIP2024]	0.0756	0.3489	0.4598	275	1673	0.0691	0.2977	0.5297	272	1445
EvDiff (Ours)	0.0463	0.3394	0.4023	148	984	0.0476	0.3677	0.4226	129	1491

which answers **①**. In the following sections, we describe how to address the scarcity of event–image paired data when training EvDiff.

3.3 Surrogate Training Pipeline

As illustrated in §3.1, there is a highly fixed degradation pattern in results from regression-based event-to-video converting methods. Hence, we use synthetic E2VID-style degradation images as a “surrogate” and propose our Surrogate Training Pipeline that enables large-scale data training, as shown in Fig. 2.

Stage 1: DiT Training. Given large-scale image dataset, we first synthesize the E2VID-style degraded images with the degradation model in §3.4. The paired images are feed into a Diffusion Transformer (DiT) for one-step diffusion training. Following the practices of diffusion-based image restoration methods [9, 85], we also make the VAE encoder (Surrogate VAE encoder) trainable to better align the latent representations with the distribution of degraded inputs. The training objective combines latent reconstruction and perceptual consistency:

$$\mathcal{L}_{\text{DiT}} = \lambda_1 \|\mathbf{z}_{\text{pred}} - \mathbf{z}_{\text{gt}}\|_2^2 + \lambda_2 \text{LPIPS}(\hat{\mathbf{I}}, \mathbf{I}_{\text{gt}}). \quad (8)$$

Stage 2: Surrogate Distillation. We then replace the current Surrogate VAE encoder with our EvEncoder, which is specifically designed to adapt to event data. As shown in Fig. 2(b), during the distillation, the event data is first converted to E2VID-style degradation images with an off-the-shelf E2VID model, and consequently encoded to $\mathbf{z}_{\text{vae}}$. As student model, EvEncoder encodes event data to $\mathbf{z}$ directly. Training follows the distillation loss:

$$\mathcal{L}_{\text{distill}} = \text{KL}(q_\phi(\mathbf{z}|E)) + \|\mathbf{z} - \mathbf{z}_{\text{vae}}\|_2^2. \quad (9)$$

Stage 3: Joint Fine-Tuning. With two main parts ready, we jointly fine-tune the whole model using small-scale event–RGB paired dataset. This stage adapts the model to genuine event distributions and enforces temporal coherence under realistic motion patterns. The joint loss integrates perceptual, flow-based temporal, and latent consistency terms:

$$\mathcal{L}_{\text{joint}} = \lambda_1 \mathcal{L}_{\text{LPIPS}} + \lambda_2 \mathcal{L}_{\text{flow}} + \lambda_3 \|\mathbf{z}_{\text{pred}} - \mathbf{z}_{\text{gt}}\|_2^2. \quad (10)$$

Although our training strategy follows a multi-stage paradigm, during inference, EvDiff predicts the reconstruction results from the event stream in a single step. Our sophisticated designed Surrogate Training Pipeline shifts the need for large-scale event–image paired data (unavailable) to widely accessible large-scale image datasets. This design enables the training of large diffusion-based foundation models for event reconstruction. The key lies in the notion of the “surrogate” — the E2VID-style degraded images, which we will introduce next.

3.4 E2VID-Style Degradation Model

We use the term “E2VID-style” to denote the common degradation characteristics observed across modern event video reconstruction models (EVRMs), including E2VID, ETNet, HyperE2VID, and related methods. It does not refer exclusively to the original E2VID architecture. Since the results from E2VID-style regression-based methods demonstrate similar degradation patterns (Fig. 3 (b)), by analyzing the factors leading to these degradations, we can synthesize low-quality (LQ) images with E2VID-style degradation from high-quality (HQ) images.

Through the whole pipeline from event to the final images, We identify three primary degradation factors in the designed E2VID-style model: event generation mechanism, sensor-inherent imperfections, and algorithmic degradation.

Event Generation Mechanism(EGM). Indicated by the first principle of event camera (Eq. 1), events only contain brightness change information, without initial condition, or absolute intensity values. This ambiguity leads to blotchy and uneven textures. To synthesize it, we first identify low-detail regions according to local variance and gradient magnitude, selecting smooth areas via percentile-based thresholding. Within these regions, irregular dark patches are stochastically generated by blending ellipses and polygons with layered noise, while multi-scale granularity enhances perceptual roughness and local contrast.

Sensor-Inherent Imperfections(SII). Due to the inherent finite pixel temporal bandwidth, trailing events persist after the actual brightness change [34]. Furthermore, random fluctuations in photon arrival and circuit leakage induce spurious events that degrade the signal-to-noise ratio of the event stream [1]. These two lead to edge degradation and noise. We synthesize them by applying soft blurring and stochastic discontinuities to locally displace boundary pixels, and adding Gaussian noise.

Algorithmic Degradation(AD). When the motion in the scene is too fast, or the accumulation time interval for events is too long, motion blur is introduced during voxelization. Moreover, event-to-image reconstruction is an under-determined inverse problem with limited spatial bandwidth. The voxel aggregation and the finite-capacity reconstruction network implicitly suppress high-frequency details and introduce spatial uncertainty.

To approximate these effects, we incorporate motion blur and a resize operation (down-sampling followed by up-sampling) into our degradation model. The resize operation simulates the spatial low-pass filtering and resolution bottleneck

Table 2: Comparison of parameters and FLOPs between our EvDiff and ControlNet at different resolutions.

Methods	Ours	ControlNet in 10 steps	ControlNet in 20 steps	ControlNet in 40 steps
#Params	2.19G	8.23G	8.23G	8.23G
Flops 512×512	2.18T	26.62T	49.93T	96.56T
Flops 1024×1024	9.56T	87.15T	163.55T	316.36T

Table 3: Ablation study on different variants of our method, with the best results highlighted in **red** and the second best in **blue**.

Method	MSE↓	SSIM↑	LPIPS↓	FID↓	FVD↓	MSE↓	SSIM↑	LPIPS↓	FID↓	FVD↓
	BSERGB					DSEC				
<i>w./o.</i> Stage 1 training	0.0726	0.2255	0.4906	374	2254	0.0630	0.3141	0.4817	256	2063
<i>w./o.</i> E2VID-Style Degradation Model	0.0850	0.1972	0.5267	396	3683	0.0918	0.2413	0.5286	450	3711
<i>w./o.</i> EGM Degradation	0.0738	0.2301	0.5000	276	1683	0.0702	0.2973	0.4937	260	1505
<i>w./o.</i> SH Degradation	0.0667	0.2478	0.4743	277	1226	0.0734	0.3017	0.4897	238	1314
<i>w./o.</i> AD Degradation	0.0747	0.223	0.5059	333	2324	0.0861	0.2774	0.5150	333	2075
<i>w./o.</i> Surrogate Distilling	0.0660	0.2822	0.4373	162	957	0.0778	0.2818	0.4750	164	1539
<i>w./o.</i> Stage 3 finetune	0.0763	0.2647	0.5740	224	1726	0.0968	0.2421	0.6546	189	1503
<i>w./o.</i> ETF	0.0586	0.2966	0.4234	159	1299	0.0605	0.3265	0.4482	152	1449
<i>w.</i> ConvGRU	0.0500	0.3302	0.4147	165	909	0.0561	0.3383	0.4337	127	985
<i>w.</i> ConvLSTM	0.0496	0.3363	0.4156	160	907	0.0620	0.3312	0.4403	127	1039
<i>w.</i> E2VID + VAE-Encoder	0.0729	0.3322	0.4827	231	1551	0.0580	0.3849	0.4691	184	1368
Final Model	0.0463	0.3394	0.4023	148	984	0.0476	0.3677	0.4226	129	1491

introduced by voxel accumulation and reconstruction, leading to detail attenuation and structural smoothing similar to those observed in E2VID outputs.

To summarize, given the HQ image, we first convert it to grayscale and apply all degradation factors with random shuffling to model E2VID-style degradation. This enables training on large-scale datasets and addresses ②.

Despite its effectiveness, the proposed framework still has several limitations:

- **Residual temporal flicker.** Our framework employs a frame-wise diffusion model and encourages temporal consistency mainly through latent consistency across frames. While this improves stability, slight illumination or appearance flicker may still occur in dynamic scenes.
- **Backbone-induced structural artifacts.** Since our implementation adopts SD3 as the generative backbone, we occasionally observe faint periodic grid-like artifacts in reconstructed frames. This is related to the patch-based architecture of the underlying DiT model and has also been reported in other SD3-based works.

4 Experiment

4.1 Experimental Settings

Training Set. We adopt the Place365 dataset [88] as our training set for stage 1, which contains 1,800,000 HQ images. For stage 2 & 3, we use the high-frame-rate videos from the REDS dataset [42], which contains 240 seq and 500 images each sequence. v2e [22] simulator is used to produce synthetic events.

Implementation Details. EvDiff is trained with the 3-stage Surrogate Training Pipeline. All models are fine-tuned using LoRA [21] with rank 64 and optimized by AdamW [36] with an initial learning rate of 5×10^{-6} ($\beta_1 = 0.9$,



Fig. 4: Visual comparison on BS-ERGB [71] and DSEC [16] datasets. Our EvDiff produces higher-quality chromatic frames.

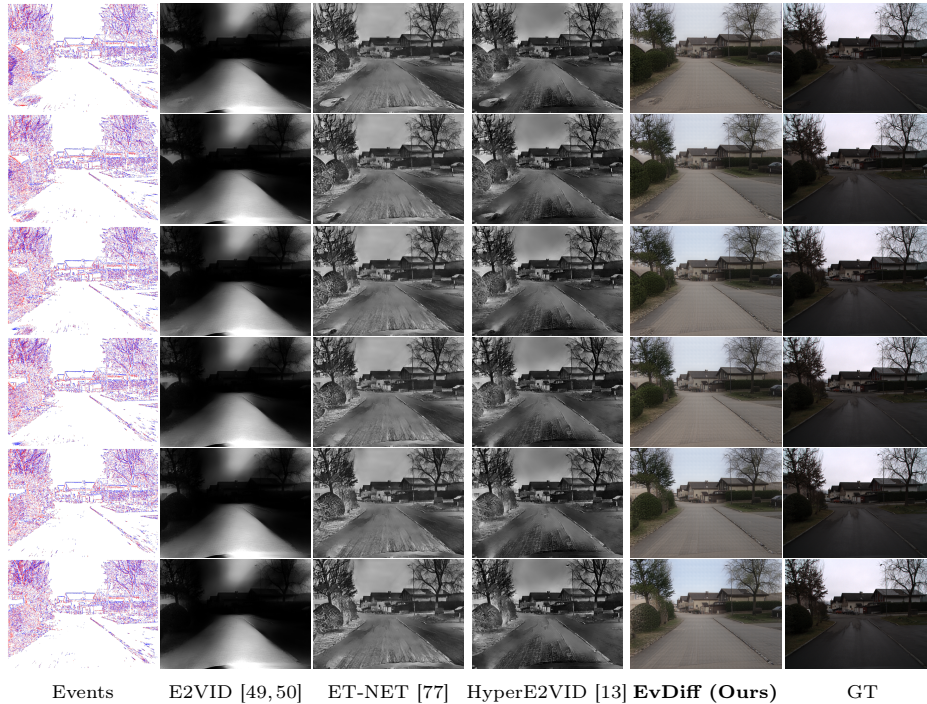


Fig. 5: Visual comparison on DSEC [16] datasets. Our EvDiff produces higher-quality chromatic frames.

$\beta_2 = 0.999$). Training is conducted on a single NVIDIA H200 GPU. Stage 1 trains the diffusion model for 180k iterations with a batch size of 10. Stage 2 performs Surrogate Distillation for 12k iterations (batch size 1, sequence length 40), followed by Stage 3 joint fine-tuning for 12k iterations (batch size 1, sequence length 30). During Stages 2 and 3, we apply online degradation to dynamically corrupt simulated event streams (e.g., event merging, dropping, and polarity removal) to better mimic real sensor behavior. More details are provided in the *supp.*

Evaluation. We evaluate our method on two real-world datasets: BS-ERGB [71] and DSEC [16] datasets. Both of them contain chromatic RGB frames and monochromatic events. We follow the official test split for DSEC and the split proposed in EVReal [12] for BS-ERGB, respectively. HQF [61] and MVSEC [89] are excluded due to their suboptimal quality and the lack of color information. We report both standard fidelity metrics (MSE, SSIM) and perceptual metrics (LPIPS, FID, FVD). For fair comparison with existing methods (*e.g.*, ETNet [77], HyperE2VID [13]), the results from our method are converted to grayscale images before computing MSE, SSIM and LPIPS metrics.

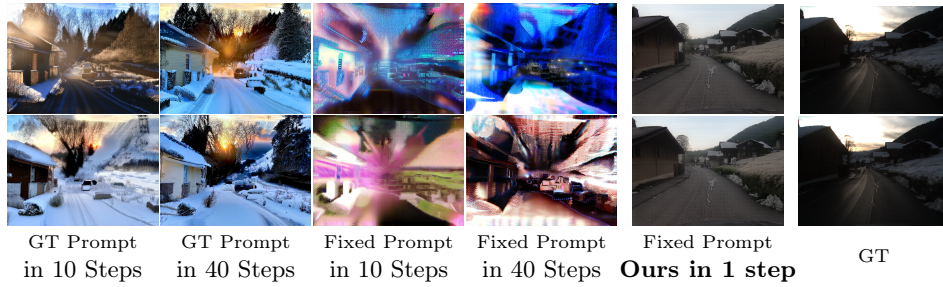


Fig. 6: Visual comparison against multi-steps ControlNet-based counterparts. Left four columns: Results from ControlNet-based methods. “GT Prompt”: Prompts are produced from GT images with RAM [79]. Our EvDiff produces more faithful results.

4.2 Comparisons with State-of-the-Art Methods

We compare our method with representative event-based video reconstruction approaches, including E2VID [49], FireNet [54], E2VID+ [61], FireNet+ [61], SPADE-E2VID [5], ETNet [77], SSL-E2VID [44], and HyperE2VID [13]. For all methods official weights are used. The numerical results on reference metrics are shown in Table 1.

Considering an improvement of 32.2%/8.5%/37.8% and 20.8%/14.0%/44.9% in MSE/LPIPS/FID compared to second best method on BS-ERGB and DSEC, respectively, the proposed EvDiff delivers overall state-of-the-art performance, showing advantages in perceptual (LPIPS) and visual realism (FID, FVD), while preserving solid fidelity (MSE, SSIM). Specifically, traditional E2VID and its variants aim to improve robustness to complex event inputs for enhancing fidelity, but their generated videos still show poor perceptual and generative quality due to the lack of priors from large dataset pretraining. This is also evident from the qualitative results in Figure 4, where the E2VID family exhibits severe visually unpleasant artifacts across the entire image, and the generated details lack realism. In contrast, benefiting from DiT training, which converts large-scale high-quality datasets into diffusion priors that can handle E2VID-style degradation artifacts, and the temporal-based EvEncoder designed for events characteristics in surrogate distillation, our EvDiff achieves significant advantages in perceptual and generative quality evaluations. Furthermore, our EvDiff is the *only* method that generates chromatic video from monochromatic events. In Fig. 4, our results retain the cloud shape, while the GT image is over-exposed. This shows that EvDiff offers higher effective dynamic range thanks to the HDR properties of event cameras.

4.3 Comparisons with the ControlNet-Based Counterpart

As described in §3.2, a more intuitive idea for using diffusion models for event-based video reconstruction is a ControlNet-based multi-step diffusion method.

We use implementation from HuggingFace [23] and keep the SD3 base model since there is no open-sourced diffusion-based E2VID method. As shown in Fig. 6, our EvDiff produces results that are more temporally consistent and more faithful to the real scene, *i.e.*, with higher *fidelity*. Moreover, when provided with the same scene-agnostic prompts (“Fixed Prompt”) used in our model, ControlNet largely fails to achieve effective reconstruction. This is crucial for video reconstruction, where events are the only input and human prior knowledge is typically unavailable. Finally, in the single-frame case, as shown in Table 2, our EvDiff contains fewer parameters and lower computational cost, not to mention that in the multi-frame setting, the computational complexity only scales linearly with the frame number T .

4.4 Ablation Study

We conduct ablation studies on both the BS-ERGB and DSEC datasets to verify the contribution of each key component, as shown in Table 3. For the proposed Surrogate Training Pipeline, removing Stage 1 leads to a 56.8% performance drop, even when initialized with pretrained SD3 weights, highlighting the necessity of large-scale dataset training. Excluding our E2VID-Style Degradation Model still results in a 83.6% performance decrease, demonstrating its importance in bridging the domain gap. We further analyze the proposed Degradation Model by ablating its internal components. When one component is excluded (either EGM, SII or AD), the performance remains inferior to the full model. Surrogate Distillation boosts performance by 29.8%, as it aligns the EvEncoder with the VAE latent space and ensures a clean, stable input for the diffusion refiner. Stage 3 joint fine-tuning is also critical. Without it, imperfections in the EvEncoder latent representations degrade the final reconstruction. Joint fine-tuning allows the DiT refiner to adapt to these imperfections and recover high-quality results.

Regarding the model architecture, the ETF module improves MSE and FVD by 20.9% and 24.2%, respectively. Replacing ETF with ConvLSTM or ConvGRU (with comparable model size) consistently degrades performance, confirming the effectiveness of the proposed temporal fusion design. Replacing our EvEncoder with the “E2VID + VAE-Encoder” baseline further degrades perceptual quality while making the pipeline $2.5\times$ larger and slower (84ms vs. 63ms), demonstrating that our EvEncoder achieves a better balance between efficiency and performance.

5 Conclusion

Event-to-video reconstruction has long been constrained by limited training data, weak generalization, and low restoration quality. By decomposing the problem and introducing the Surrogate Training Pipeline, EvDiff transfers the power of large pretrained generative models to event-based video reconstruction and enables training with large-scale image datasets. The resulting one-step

EvDiff produces high-quality, photorealistic, and faithful reconstructions from event streams while remaining competitive with ControlNet-based approaches. We hope this work helps bridge event-based vision and modern generative models and inspires future research on temporally consistent event-to-video reconstruction.

Acknowledgements

This research was supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LZ24F050003. It was also partially funded by the Ministry of Education and Science of Bulgaria, as part of the Bulgarian National Roadmap for Research Infrastructure (support for INSAIT) and the European Union’s Horizon Europe – the Framework Programme for Research and Innovation, under grant agreement 101168521.

References

1. iniVation A.G.: Understanding the performance of neuromorphic event-based vision sensors. Tech. Rep. (2020)
2. Bao, Y., Sun, L., Ma, Y., Gu, D., Wang, K.: Improving fast auto-focus with event polarity. *Optics Express* **31**(15), 24025–24044 (2023)
3. Bardow, P., Davison, A.J., Leutenegger, S.: Simultaneous optical flow and intensity estimation from an event camera. In: *Proc. CVPR* (2016)
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127* (2023)
5. Cadena, P.R.G., Qian, Y., Wang, C., Yang, M.: Spade-e2vid: Spatially-adaptive denormalization for event-based video reconstruction. *IEEE Transactions on Image Processing* **30**, 2488–2500 (2021). <https://doi.org/10.1109/TIP.2021.3052070>
6. Cadena, P.R.G., Qian, Y., Wang, C., Yang, M.: Sparse-e2vid: A sparse convolutional model for event-based video reconstruction trained with real event noise. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4150–4158 (2023)
7. Chen, H., Teng, M., Shi, B., Wang, Y., Huang, T.: Learning to deblur and generate high frame rate video with an event camera. *arXiv preprint arXiv:2003.00847* (2020)
8. Chen, J., Shi, H., Ye, Y., Yang, K., Sun, L., Wang, K.: Efficient human pose estimation via 3d event point cloud. In: *2022 International Conference on 3D Vision (3DV)*. pp. 1–10. IEEE (2022)
9. Chen, J., Pan, J., Dong, J.: Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In: *CVPR* (2025)
10. Cho, H., Kim, T., Jeong, Y., Yoon, K.J.: A benchmark dataset for event-guided human pose estimation and tracking in extreme conditions. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) *Advances in Neural Information Processing Systems*. vol. 37, pp. 134826–134840. Curran Associates, Inc. (2024). <https://doi.org/10.52202/>

- 079017-4284, https://proceedings.neurips.cc/paper_files/paper/2024/file/f304e427cfe6bb762fe1bf18516c8a87-Paper-Datasets_and_Benchmarks_Track.pdf, accessed: 2026-06-25
11. Dong, L., Fan, Q., Guo, Y., Wang, Z., Zhang, Q., Chen, J., Luo, Y., Zou, C.: Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23174–23184 (2025)
 12. Ercan, B., Eker, O., Erdem, A., Erdem, E.: Evreal: Towards a comprehensive benchmark and analysis suite for event-based video reconstruction. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3943–3952 (2023). <https://doi.org/10.1109/CVPRW59228.2023.00410>
 13. Ercan, B., Eker, O., Saglam, C., Erdem, A., Erdem, E.: HyperE2VID: Improving event-based video reconstruction via hypernetworks. IEEE Transactions on Image Processing **33**, 1826–1837 (2024). <https://doi.org/10.1109/TIP.2024.3372460>
 14. Esser, P., Kulal, S., Blattmann, A., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of the 41st International Conference on Machine Learning (2024)
 15. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(1), 154–180 (2020)
 16. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. IEEE Robotics and Automation Letters (2021). <https://doi.org/10.1109/LRA.2021.3068942>
 17. Han, J., Yang, Y., Zhou, C., Xu, C., Shi, B.: EvIntSR-net: Event guided multiple latent frames reconstruction and super-resolution. In: Proc. ICCV (2021)
 18. He, W., You, K., Qiao, Z., Jia, X., Zhang, Z., Wang, W., Lu, H., Wang, Y., Liao, J.: Timereplayer: Unlocking the potential of event cameras for video interpolation. In: Proc. CVPR. pp. 17804–17813 (2022)
 19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
 20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
 21. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
 22. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic dvs events. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1312–1321 (2021)
 23. Hugging Face, Inc.: Controlnet with stable diffusion 3 — diffusers documentation. https://huggingface.co/docs/diffusers/main/en/api/pipelines/controlnet_sd3 (2024), accessed: 2026-06-25
 24. Jiang, Z., Zhang, Y., Zou, D., Ren, J., Lv, J., Liu, Y.: Learning event-based motion deblurring. In: Proc. CVPR. pp. 3320–3329 (2020)
 25. Kim, T., Chae, Y., Jang, H.K., Yoon, K.J.: Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18032–18042 (2023). <https://doi.org/10.1109/CVPR52729.2023.01729>
 26. Kim, T., Cho, H., Yoon, K.J.: Frequency-aware event-based video deblurring for real-world motion blur. In: 2024 IEEE/CVF Conference on Computer Vision and

- Pattern Recognition (CVPR). pp. 24966–24976 (2024). <https://doi.org/10.1109/CVPR52733.2024.02358>
27. Kim, T., Cho, H., Yoon, K.J.: Cmta: Cross-modal temporal alignment for event-guided video deblurring. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) *Computer Vision – ECCV 2024*. pp. 1–19. Springer Nature Switzerland, Cham (2025)
 28. Kim, T., Jeong, J., Cho, H., Jeong, Y., Yoon, K.J.: Towards real-world event-guided low-light video enhancement and deblurring. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) *Computer Vision – ECCV 2024*. pp. 433–451. Springer Nature Switzerland, Cham (2025)
 29. Kim, T., Lee, J., Wang, L., Yoon, K.J.: Event-guided deblurring of unknown exposure time videos. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) *Computer Vision – ECCV 2022*. pp. 519–538. Springer Nature Switzerland, Cham (2022)
 30. Kim, T.K., Yoon, K.J.: Event-guided unified framework for low-light video enhancement, frame interpolation, and deblurring. In: *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 8524–8534 (2025). <https://doi.org/10.1109/ICCV51701.2025.00798>
 31. Liang, J., Yu, B., Yang, Y., Han, Y., Shi, B.: E2vidiff: Perceptual events-to-video reconstruction using diffusion priors. *arXiv preprint arXiv:2407.08231* (2024)
 32. Lin, S., Xia, X., Ren, Y., Yang, C., Xiao, X., Jiang, L.: Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316* (2025)
 33. Lin, S., Zhang, J., Pan, J., Jiang, Z., Zou, D., Wang, Y., Chen, J., Ren, J.: Learning event-driven video deblurring and interpolation. In: *Proc. ECCV*. pp. 695–710. Springer (2020)
 34. Liu, H., Peng, S., Zhu, L., Chang, Y., Zhou, H., Yan, L.: Seeing motion at nighttime with an event camera. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 25648–25658 (2024)
 35. Liu, X., Wang, Y., Chen, Z., Cao, J., Zhang, H., Zhang, Y., Yang, X.: One-step diffusion model for image motion-deblurring. *arXiv preprint arXiv:2503.06537* (2025)
 36. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net (2019), <https://openreview.net/forum?id=Bkg6RiCqY7>, accessed: 2026-06-25
 37. Lu, X., Sun, L., Gu, D., Wang, K.: Sge: structured light system based on gray code with an event camera. *Optics Express* **32**(26), 46044–46061 (2024)
 38. Lu, Y., Shi, D., Li, R., Zhang, Y., Jing, L., Yang, S.: Scse-e2vid: Improved event-based video reconstruction with an event camera. In: *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. pp. 3249–3254 (2022). <https://doi.org/10.1109/SMC53654.2022.9945237>
 39. Mahowald, M.: The silicon retina. In: *An Analog VLSI System for Stereoscopic Vision*, pp. 4–65. Springer (1994)
 40. Manderscheid, J., Sironi, A., Bourdis, N., Migliore, D., Lepetit, V.: Speed invariant time surface for learning to detect corner points with event-based cameras. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10245–10254 (2019)
 41. Munda, G., Reinbacher, C., Pock, T.: Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision* **126**(12), 1381–1393 (2018)

42. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
43. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a blurry frame alive at high frame-rate with an event camera. In: Proc. CVPR. pp. 6820–6829 (2019)
44. Paredes-Vallés, F., de Croon, G.C.H.E.: Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3445–3454 (2021). <https://doi.org/10.1109/CVPR46437.2021.00345>
45. Patrick, L., Posch, C., Delbruck, T.: A 128×128 120 dB 15μ s latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits (2008)
46. Qian, X., Jiang, Q., Sun, L., Yu, Z., Yang, K., Wu, P., Zhou, J., Gao, Y., Ma, Y., Yang, M.H., et al.: Learning latent transmission and glare maps for lens veiling glare removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 33995–34005 (2026)
47. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)
48. Rebecq, H., Gehrig, D., Scaramuzza, D.: Esim: an open event camera simulator. In: Conference on robot learning. pp. 969–982. PMLR (2018)
49. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3857–3866 (2019)
50. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE transactions on pattern analysis and machine intelligence **43**(6), 1964–1980 (2019)
51. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
52. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
53. Scheerlinck, C., Barnes, N., Mahony, R.: Continuous-time intensity estimation using event cameras. In: Asian Conference on Computer Vision. pp. 308–324. Springer (2018)
54. Scheerlinck, C., Rebecq, H., Gehrig, D., Barnes, N., Mahony, R.E., Scaramuzza, D.: Fast image reconstruction with an event camera. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 156–163 (2020). <https://doi.org/10.1109/WACV45572.2020.9093366>
55. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)
56. SHI, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.k., WOO, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates,

- Inc. (2015), https://proceedings.neurips.cc/paper_files/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf, accessed: 2026-06-25
57. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
 58. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), <https://openreview.net/forum?id=St1giarCHLP>, accessed: 2026-06-25
 59. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **32** (2019)
 60. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), <https://openreview.net/forum?id=PXTIG12RRHS>, accessed: 2026-06-25
 61. Stoffregen, T., Scheerlinck, C., Scaramuzza, D., Drummond, T., Barnes, N., Klee-man, L., Mahony, R.: Reducing the sim-to-real gap for event cameras. In: European Conference on Computer Vision. pp. 534–549. Springer (2020)
 62. Su, S., Rocafort, J.M., Xue, D., Serrano-Lozano, D., Sun, L., Vazquez-Corral, J.: Bridging the perception gap in image super-resolution evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 30532–30542 (2026)
 63. Sun, L., Alfarano, A., Duan, P., Su, S., Wang, K., Shi, B., Timofte, R., Paudel, D.P., Van Gool, L., Liu, Q., et al.: Ntire 2025 challenge on event-based image deblurring: Methods and results. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1324–1341 (2025)
 64. Sun, L., Bao, Y., Zhai, J., Liang, J., Zhang, Y., Wang, K., Paudel, D.P., Van Gool, L.: Low-light image enhancement using event-based illumination estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6667–6677 (2025)
 65. Sun, L., Gehrig, D., Sakaridis, C., Gehrig, M., Liang, J., Sun, P., Xu, Z., Wang, K., Van Gool, L., Scaramuzza, D.: A unified framework for event-based frame interpolation with ad-hoc deblurring in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)
 66. Sun, L., Li, W., Wang, X., Li, Z., Shi, L., Xu, D., Zhang, D., Hu, M., Guo, S., Su, S., et al.: The second challenge on event-based image deblurring at ntire 2026: Methods and results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2544–2561 (2026)
 67. Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., Gool, L.V.: Event-based fusion for motion deblurring with cross-modal attention. In: Proc. ECCV. pp. 412–428. Springer (2022)
 68. Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., Van Gool, L.: Mefnet: Multi-scale event fusion network for motion deblurring. *arXiv preprint arXiv:2112.00167* (2021)
 69. Sun, L., Sakaridis, C., Liang, J., Sun, P., Cao, J., Zhang, K., Jiang, Q., Wang, K., Van Gool, L.: Event-based frame interpolation with ad-hoc deblurring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 22871. IEEE (2023)
 70. Tavanaei, A., Ghodrati, M., Kheradpisheh, S.R., Masquelier, T., Maida, A.: Deep learning in spiking neural networks. *Neural networks* **111**, 47–63 (2019)

71. Tulyakov, S., Boicchio, A., Gehrig, D., Georgoulis, S., Li, Y., Scaramuzza, D.: Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In: Proc. CVPR. pp. 17755–17764 (2022)
72. Tulyakov, S., Gehrig, D., Georgoulis, S., Erbach, J., Gehrig, M., Li, Y., Scaramuzza, D.: Time lens: Event-based video frame interpolation. In: Proc. CVPR. pp. 16155–16164 (2021)
73. Vitoria, P., Georgoulis, S., Tulyakov, S., Boicchio, A., Erbach, J., Li, Y.: Event-based image deblurring with dynamic motion awareness. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) Computer Vision – ECCV 2022 Workshops. pp. 95–112. Springer Nature Switzerland, Cham (2023)
74. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
75. Wang, L., Ho, Y.S., Yoon, K.J., et al.: Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In: Proc. CVPR. pp. 10081–10090 (2019)
76. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision **133**(5), 3059–3078 (2025)
77. Weng, W., Zhang, Y., Xiong, Z.: Event-based video reconstruction using transformer. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2543–2552 (2021). <https://doi.org/10.1109/ICCV48922.2021.00256>
78. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. Advances in Neural Information Processing Systems **37**, 92529–92553 (2024)
79. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: Seesr: Towards semantics-aware real-world image super-resolution. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25456–25467 (2024). <https://doi.org/10.1109/CVPR52733.2024.02405>
80. Xu, F., Yu, L., Wang, B., Yang, W., Xia, G.S., Jia, X., Qiao, Z., Liu, J.: Motion deblurring with real events. In: Proc. ICCV (2021)
81. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications. ACM computing surveys **56**(4), 1–39 (2023)
82. Ye, Y., Shi, H., Yang, K., Wang, Z., Yin, X., Sun, L., Wang, Y., Wang, K.: Towards anytime optical flow estimation with event cameras. Sensors **25**(10), 3158 (2025)
83. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024)
84. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
85. Yu, F., Gu, J., Li, Z., Hu, J., Kong, X., Wang, X., He, J., Qiao, Y., Dong, C.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild (2024)
86. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
87. Zhang, X., Yu, L.: Unifying motion deblurring and frame interpolation with events. In: Proc. CVPR. pp. 17765–17774 (2022)

- 88. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017)
- 89. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters* **3**(3), 2032–2039 (2018)
- 90. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898* (2018)