

# EvDiff: High Quality Video with an Event Camera

Weilun Li<sup>1,2\*</sup> Lei Sun<sup>2\*†</sup> Ruixi Gao<sup>1</sup> Qi Jiang<sup>1</sup> Yuqin Ma<sup>1</sup> Kaiwei Wang<sup>1†</sup> Ming-Hsuan Yang<sup>3,4</sup>  
 Luc Van Gool<sup>2</sup> Danda Pani Paudel<sup>2</sup>  
<sup>1</sup>Zhejiang University      <sup>2</sup>INSAIT      <sup>3</sup>UC Merced      <sup>4</sup>Google DeepMind

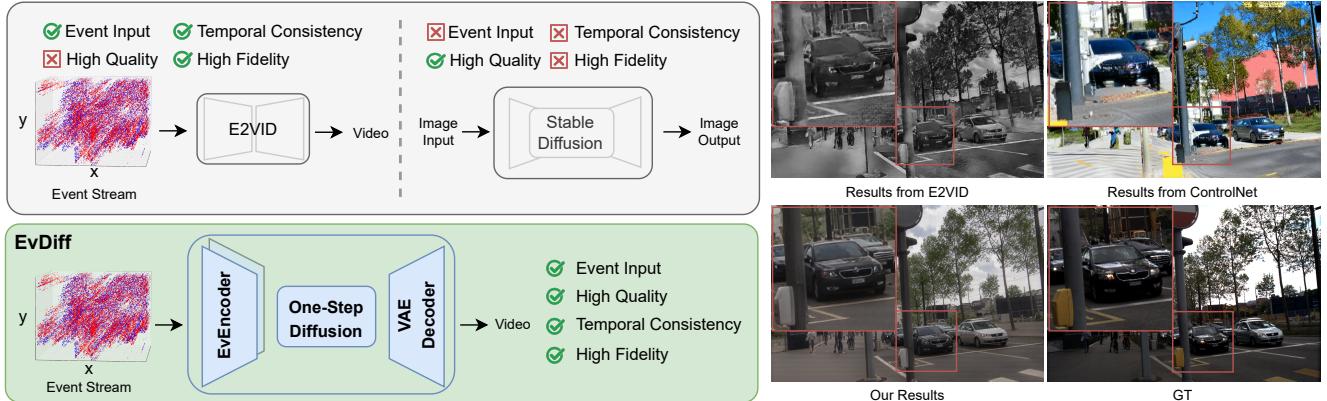


Figure 1. Our EvDiff can reconstruct real high-quality video streams from monochrome event streams, while maintaining both fidelity and realism. Compared with Ground-Truth (GT), our result shows a higher dynamic range.

## 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 heavy 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 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 upon acceptance.

## 1. Introduction

Inspired by the human visual system, the Silicon Retina [31] 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 [12, 38].

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) [63] are employed. To mitigate this gap, researchers often convert event streams

\*Work done during Weilun Li's internship at INSAIT.

†Equal contribution. †Corresponding authors.

into middle representations, such as voxel grids [83] or time surfaces [32]. 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 [41, 42].

While early works [2, 33, 46] incorporated handcrafted smoothness priors into video reconstruction, a major breakthrough was achieved by E2VID [41], which introduced a data-driven learning approach to event-to-video conversion. Subsequent studies have refined this paradigm by simplifying the network architecture [47], adding self-supervised constraints [36], and explicitly modeling the sparsity of events [5], 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 [16, 44, 75]. 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 [49, 81].

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: ① *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. ② 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 [19, 40] also require high-frame-rate videos as inputs, which are similarly 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 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 spe-

cific, 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 [81], 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 [79], 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 present the first approach 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 [22, 23, 58, 59, 62, 66, 73], and video interpolation [6, 15, 25, 35, 60, 61, 64, 65, 80]. In addition to these tasks, we investigate the use of event cameras for video reconstruction [1, 2, 5, 14, 33, 36, 41, 42, 46, 47, 68], *i.e.*, recovering intensity frames from event streams, in this work.

Early approaches [2, 33, 46] 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 sig-

nificantly. A milestone was E2VID [41, 42], which combined a U-Net with temporal modules and achieved strong results at moderate computational cost. Later works such as FireNet [48] explored lightweight architectures, while E2VID+ and FireNet+ [56] addressed the shortage of training data by introducing synthetic events generated with ESIM [40], a strategy widely adopted in subsequent research. Following E2VID-style variants (*e.g.*, SPADE-E2VID [4], Hyper-E2VID [10], SCSE-E2VID [30], Sparse-E2VID [5]) 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 [36] 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. More recently, endeavors have sought to move beyond this paradigm. For example, E2VIDiff [24] introduces a diffusion-based approach that conditions Stable Diffusion [43] 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 [16, 52, 54, 55], 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 [74]. When scaled up with billions of parameters and trained on massive natural image datasets [49, 69], they demonstrate impressive realism and strong generative priors, enabling high-quality image and video synthesis [3, 39, 43, 45, 67]. However, these benefits come at the cost of heavy computational requirements, as classical diffusion processes involve hundreds of iterative denoising steps [16, 53], 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 [26, 76, 77]. 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 [8, 71], image deblurring [28], *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 [41, 42], 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 [33, 46], 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 strong 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.

### 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:

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

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

where  $\mathcal{D}$  is the VAE decoder, and  $\{\mathbf{z}_i\}_{i=1}^N$  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  $\Delta t$ , we first encode them into latent representation. To be specific, asynchronous event streams are first converted to synchronous voxel grids [83]  $\mathbf{V} \in \mathbb{R}^{H \times W \times T}$  which contain temporal information in  $c$  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 [79] for self-attention, similar to SVD [3], 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 [42], which substantially increases the temporal length  $T$  and consequently leads to unacceptable computational overhead.

Therefore, we design a recurrent architecture [17, 51] in EvEncoder to efficiently keep the temporal information flow. To maintain efficiency, we propose the Efficient Temporal Fusion (ETF) module in EvEncoder, which employs a lightweight gated fusion mechanism [50], as illustrated in Fig. 2 (b). Each ETF module incorporates 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 ETF module, we reduce the computation 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 [8, 71] 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) [11] 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:

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

where  $\epsilon$  denotes a denoising network,  $\alpha_{t^*}$  and  $\beta_{t^*}$  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, 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 [7, 78], 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  $\mathbf{z}_{vae}$ . As

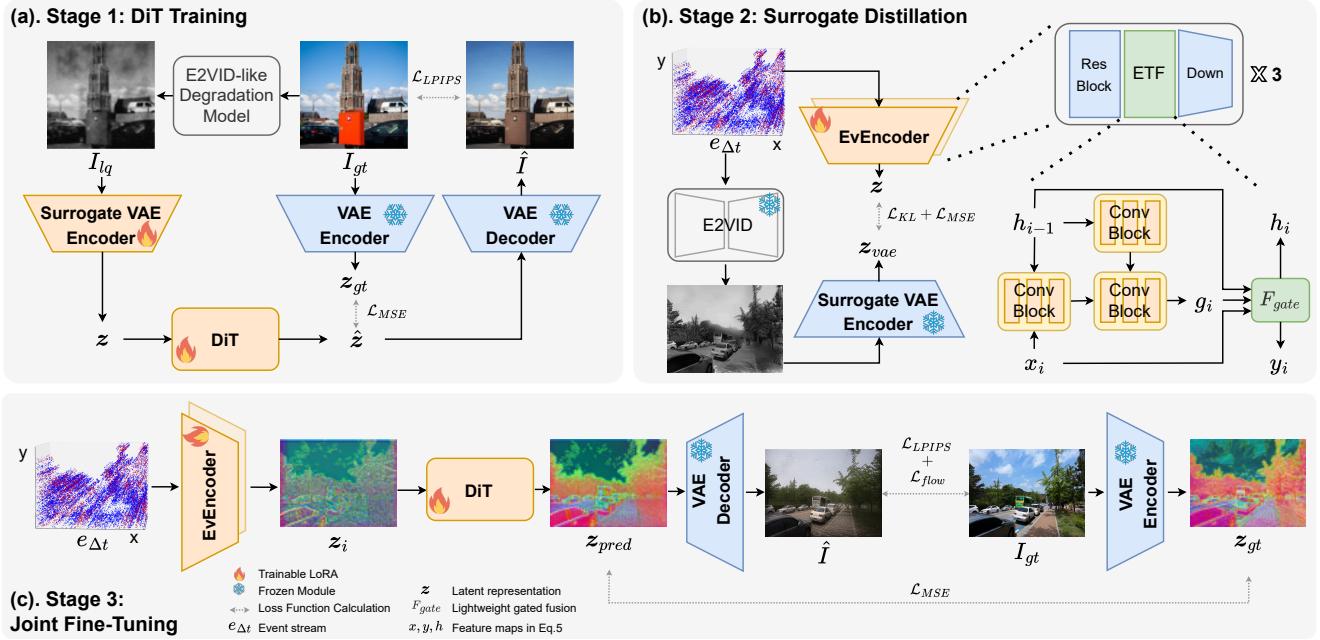


Figure 2. **The proposed Surrogate Training Pipeline.** Stage 1: We train a DiT model and Surrogate VAE Encoder with LQ-HQ pairs; Stage 2: The Surrogate VAE Encoder is distilled into EvEncoder; Stage 3: We finetune the whole EvDiff model.

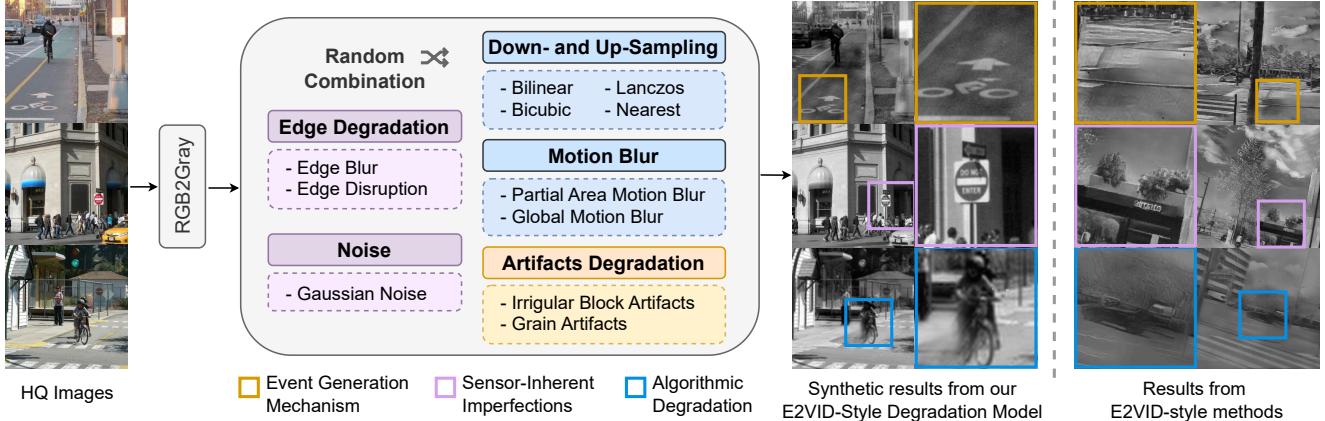


Figure 3. **Overview of the proposed E2VID-Style Degradation Model.** HQ images sampled from the Places365 [81] dataset are cropped to  $512 \times 512$  and then processed by our degradation model to generate corresponding LQ images. For comparison, E2VID-style results from the BS-ERGB dataset are shown on the right.

student model, EvEncoder encodes event data to  $z$  directly. Training follows the distillation loss:

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

**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 inte-

grates perceptual, flow-based temporal, and latent consistency terms:

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

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 (which are unavailable) to widely accessible large-scale image datasets. This design enables the training of large diffusion-based

Table 1. **Quantitative comparison** on BS-ERGB [65] and DSEC [13], with the best results highlighted in red and the second best in blue.

Method	MSE↓	SSIM↑	LPIPS↓	FID↓	FVD↓	MSE↓	SSIM↑	LPIPS↓	FID↓	FVD↓
<b>BS-ERGB</b>					<b>DSEC</b>					
E2VID <sub>[CVPR2019]</sub> [41, 42]	0.1175	0.3313	0.5518	276	1688	0.1042	0.3393	0.5427	239	1541
FireNet <sub>[WACV2020]</sub> [48]	0.0857	0.3302	0.5328	271	1909	0.0927	0.3478	0.5670	262	1492
E2VID+ <sub>[ECCV2020]</sub> [56]	0.0717	0.3710	0.4397	238	1652	0.0881	0.2812	0.5138	248	1397
FireNet+ <sub>[ECCV2020]</sub> [56]	0.0827	0.3080	0.4903	313	1562	0.0913	0.2268	0.5655	291	1592
SPADE-E2VID <sub>[TIP2021]</sub> [4]	0.0891	0.3235	0.6715	347	2343	0.0601	0.4578	0.4911	234	1277
SSL-E2VID <sub>[CVPR2021]</sub> [37]	0.0798	0.3448	0.6169	312	1907	0.1055	0.3303	0.5565	326	1521
ET-Net <sub>[ICCV2021]</sub> [70]	0.0683	0.3557	0.4566	270	1569	0.0737	0.2953	0.5232	259	1496
HyperE2VID <sub>[TIP2024]</sub> [10]	0.0756	0.3489	0.4598	275	1673	0.0691	0.2977	0.5297	272	1445
<b>EvDiff (Ours)</b>	<b>0.0463</b>	0.3394	<b>0.4023</b>	<b>148</b>	<b>984</b>	<b>0.0476</b>	<b>0.3677</b>	<b>0.4226</b>	<b>129</b>	1491

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

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.** 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 elliptical and polygonal bases with layered noise, while multi-scale granularity enhances perceptual roughness and local contrast.

**Sensor-Inherent Imperfections.** Due to the inherent finite pixel temporal bandwidth, trailing events persist after the actual brightness change [27]. Furthermore, random fluctuations in photon arrival and circuit leakage induce spurious events that degrade the signal-to-noise ratio of the event stream [21]. 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.** When the motion in the scene is too fast, or the accumulation time interval for events is too long, there is also motion blur when we convert events to voxel grids. Besides, the Unet-like architecture in E2VID-style methods also introduces losses in the spatial information. Therefore, we add motion blur and resize (Down- and up-sampling) operation in our degradation model.

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	2.18T	26.62T	49.93T	96.56T
512×512				
Flops	9.56T	87.15T	163.55T	316.36T
1024×1024				

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

## 4. Experiment

### 4.1. Experimental Settings

**Training Set.** We adopt the Place365 dataset [81] 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 [34], which contains 240 seq and 500 images each sequence. v2e [19] simulator is used to produce synthetic events.

**Implementation Details.** Our EvDiff follows a 3-stage Surrogate Training Pipeline. All models are fine-tuned using LoRA [18] with a rank of 64, optimized by AdamW [29] with an initial learning rate of  $5 \times 10^{-6}$ ,  $\beta_1 = 0.9$ , and  $\beta_2 = 0.999$ . All training is conducted on a single NVIDIA H200 GPU. For stage 1, we train the diffusion model for 180,000 iterations with a batch size of 10. In stage 2, the Surrogate Distillation is then trained for 12,000 iterations with a batch size of 1 and a sequence length of 40. For stage 3, the whole model is fine-tuned for 12,000 iterations using a batch size of 1 and a sequence length of 30. For the training in both Stage 2 & 3, we applied online degradation to dynamically corrupt the simulated event streams, including random merging of adjacent events, probabilistic event dropping, and localized polarity removal. which dynamic corruption better reflects real sensor behavior. For more details, kindly refer to *supp*.

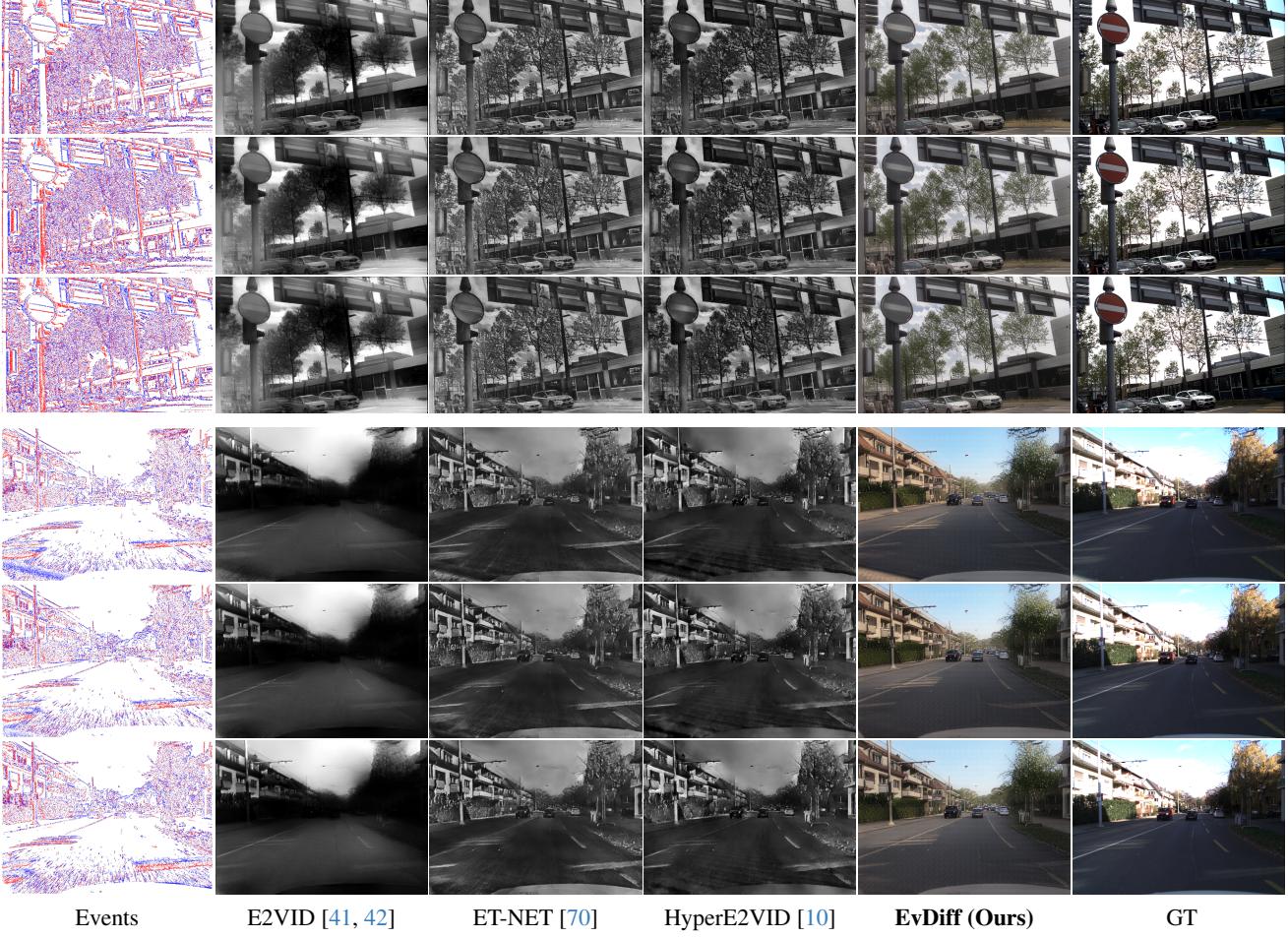


Figure 4. **Visual comparison on BS-ERGB [65] and DSEC [13] datasets.** Our EvDiff produces higher-quality chromatic frames.

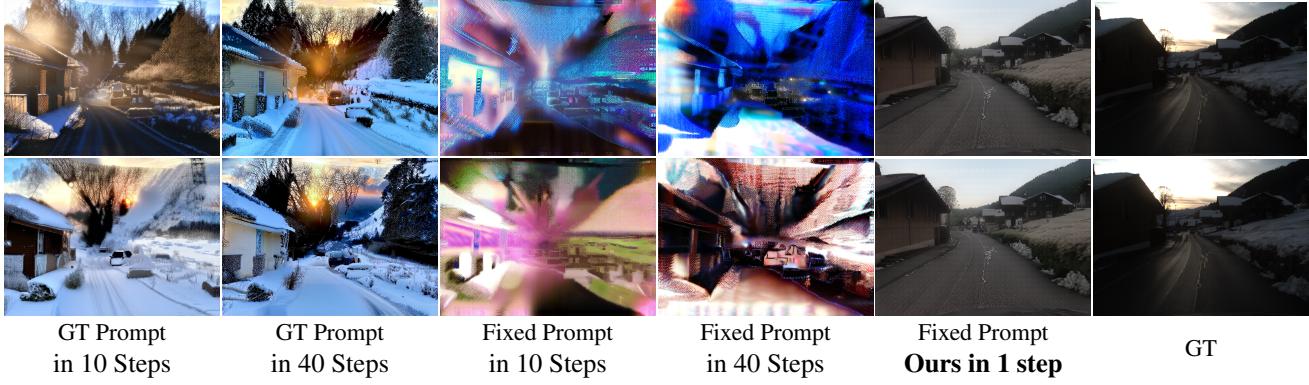


Figure 5. **Visual comparison against multi-steps ControlNet-based counterparts.** Left four columns: Results from ControlNet-based methods. “GT Prompt”: Prompts are produced from GT images with RAM [72]. Our EvDiff produces more faithful results.

**Evaluation.** We evaluate our method on two real-world datasets: BS-ERGB [65] and DSEC [13] 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 [9] for BS-ERGB, respectively. HQF [57] and MVSEC [82] are excluded due to their sub-

optimal 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 [70], HyperE2VID [10]), the results from our method are converted to grayscale images before computing MSE, SSIM and LPIPS metrics.

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				
w./o. Stage 1 training	0.0726	0.2255	0.4906	374	2254	0.0630	0.3141	0.4817	256	2063
w./o. E2VID-Style Degradation Model	<b>0.0525</b>	0.3224	<b>0.4193</b>	<b>148</b>	1006	0.0671	0.3148	0.4585	<b>138</b>	<b>1413</b>
w./o. Surrogate Distilling	0.0660	0.2822	0.4373	162	<b>957</b>	0.0778	0.2818	0.4750	164	1539
w./o. ETF	0.0586	0.2966	0.4234	159	1299	0.0605	0.3265	<b>0.4482</b>	152	1449
w. E2VID + VAE-Encoder	0.0729	<b>0.3322</b>	0.4827	231	1551	<b>0.0580</b>	<b>0.3849</b>	0.4691	184	<b>1368</b>
Final Model	<b>0.0463</b>	<b>0.3394</b>	<b>0.4023</b>	<b>148</b>	<b>984</b>	<b>0.0476</b>	<b>0.3677</b>	<b>0.4226</b>	<b>129</b>	1491

## 4.2. Comparisons with State-of-the-Art Methods

We compare our method with representative event-based video reconstruction approaches, including E2VID [41], FireNet [48], E2VID+ [56], FireNet+ [56], SPADE-E2VID [4], ETNet [70], SSL-E2VID [37], and HyperE2VID [10]. 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% (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 [20] and keep the SD3 base model since there is no open-sourced diffusion-based E2VID method. As shown in Fig. 5, our EvDiff produces results that are more temporally consistent and more faithful to the real scene, *i.e.*, with higher *fidelity*. More-

over, 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% 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 39% performance decrease, demonstrating its importance in bridging the domain gap. Finally, 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. Regarding the model architecture, the ETF module improves MSE and FVD by 20.9% and 24.2%, resp., indicating better reconstruction fidelity and temporal consistency. Moreover, replacing our EvEncoder with the “E2VID + VAE-Encoder” baseline leads to degraded perceptual performance while making the model 2.5× larger and much slower (84 ms vs. 63 ms), which shows that our EvEncoder achieves a 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 dividing the task into two parts and using the Surrogate Training Pipeline, our EvDiff effectively transfers the representational power of large pretrained models and leverages large-scale image data for training. The efficient one-step EvDiff enables high-quality, photorealistic, and faithful video reconstruction, competing favorably against the ControlNet counterpart. We hope this work sheds light on event-based video reconstruction and, more broadly, contributes to the event-based vision community.

## References

- [1] Yuhang Bao, Lei Sun, Yuqin Ma, and Kaiwei Wang. Temporal-mapping photography for event cameras. *arXiv preprint arXiv:2403.06443*, 2024. 2
- [2] Patrick Bardow, Andrew J. Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proc. CVPR*, 2016. 2
- [3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3, 4
- [4] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. Spade-e2vid: Spatially-adaptive de-normalization for event-based video reconstruction. *IEEE Transactions on Image Processing*, 30:2488–2500, 2021. 3, 6, 8
- [5] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. 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*, pages 4150–4158, 2023. 2, 3
- [6] Haoyu Chen, Minggui Teng, Boxin Shi, Yizhou Wang, and Tiejun Huang. Learning to deblur and generate high frame rate video with an event camera. *arXiv preprint arXiv:2003.00847*, 2020. 2
- [7] Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *CVPR*, 2025. 4
- [8] Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. 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*, pages 23174–23184, 2025. 3, 4
- [9] Burak Ercan, Onur Eker, Aykut Erdem, and Erkut Erdem. 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)*, pages 3943–3952, 2023. 7
- [10] 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. 3, 6, 7, 8
- [11] Patrick Esser, Sameer Kulal, Andreas Blattmann, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 4
- [12] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2020. 1
- [13] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 2021. 6, 7
- [14] Jin Han, Yixin Yang, Chu Zhou, Chao Xu, and Boxin Shi. EvIntSR-net: Event guided multiple latent frames reconstruction and super-resolution. In *Proc. ICCV*, 2021. 2
- [15] Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, and Jianxing Liao. Timereplayer: Unlocking the potential of event cameras for video interpolation. In *Proc. CVPR*, pages 17804–17813, 2022. 2
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3
- [17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4
- [18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6
- [19] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbrück. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1312–1321, 2021. 2, 6
- [20] Hugging Face, Inc. Controlnet with stable diffusion 3 — diffusers documentation. [https://huggingface.co/docs/diffusers/main/en/api/pipelines/controlnet\\_sd3](https://huggingface.co/docs/diffusers/main/en/api/pipelines/controlnet_sd3), 2024. Accessed: 2025-11-13. 8
- [21] iniVation A G. Understanding the performance of neuromorphic event-based vision sensors. *Tech. Rep.*, 2020. 6
- [22] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *Proc. CVPR*, pages 3320–3329, 2020. 2
- [23] Taewoo Kim, Jungmin Lee, Lin Wang, and Kuk-Jin Yoon. Event-guided deblurring of unknown exposure time videos. *arXiv preprint arXiv:2112.06988*, 2021. 2
- [24] Jinxiu Liang, Bohan Yu, Yixin Yang, Yiming Han, and Boxin Shi. E2vidiff: Perceptual events-to-video reconstruction using diffusion priors. *arXiv preprint arXiv:2407.08231*, 2024. 3
- [25] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *Proc. ECCV*, pages 695–710. Springer, 2020. 2
- [26] Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025. 3
- [27] Haoyue Liu, Shihan Peng, Lin Zhu, Yi Chang, Hanyu Zhou, and Luxin Yan. Seeing motion at nighttime with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25648–25658, 2024. 6
- [28] Xiaoyang Liu, Yuquan Wang, Zheng Chen, Jiezhang Cao, He Zhang, Yulun Zhang, and Xiaokang Yang. One-step diffusion model for image motion-deblurring. *arXiv preprint arXiv:2503.06537*, 2025. 3
- [29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

- [30] Yue Lu, Dianxi Shi, Ruihao Li, Yi Zhang, Luoxi Jing, and Shaowu Yang. Scse-e2vid: Improved event-based video reconstruction with an event camera. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3249–3254, 2022. 3
- [31] Misha Mahowald. The silicon retina. In *An Analog VLSI System for Stereoscopic Vision*, pages 4–65. Springer, 1994. 1
- [32] Jacques Manderscheid, Amos Sironi, Nicolas Bourdis, Davide Migliore, and Vincent Lepetit. 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*, pages 10245–10254, 2019. 2
- [33] 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, 3
- [34] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. 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*, pages 0–0, 2019. 6
- [35] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proc. CVPR*, pages 6820–6829, 2019. 2
- [36] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *Proc. CVPR*, pages 3446–3455, 2021. 2, 3
- [37] 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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3445–3454, 2021. 6, 8
- [38] Lichtsteiner Patrick, Christoph Posch, and Tobi Delbrück. A 128×128 120 dB 15μ s latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 2008. 1
- [39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 3
- [40] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on robot learning*, pages 969–982. PMLR, 2018. 2, 3
- [41] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019. 2, 3, 6, 7, 8
- [42] 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, 2019. 2, 3, 4, 6, 7
- [43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3
- [46] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision*, pages 308–324. Springer, 2018. 2, 3
- [47] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pages 156–163, 2020. 2
- [48] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert E. Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 156–163, 2020. 3, 6, 8
- [49] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, 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. 2, 3
- [50] 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. 2015. 4
- [51] 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. *Advances in neural information processing systems*, 28, 2015. 4
- [52] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 3
- [53] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3
- [54] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 3
- [55] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

- [56] 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 *European Conference on Computer Vision*, pages 534–549. Springer, 2020. 3, 6, 8
- [57] 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 *European Conference on Computer Vision*, pages 534–549. Springer, 2020. 7
- [58] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Mefnet: Multi-scale event fusion network for motion deblurring. *arXiv preprint arXiv:2112.00167*, 2021. 2
- [59] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *Proc. ECCV*, pages 412–428. Springer, 2022. 2
- [60] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 22871. IEEE, 2023. 2
- [61] Lei Sun, Daniel Gehrig, Christos Sakaridis, Mathias Gehrig, Jingyun Liang, Peng Sun, Zhiping Xu, Kaiwei Wang, Luc Van Gool, and Davide Scaramuzza. A unified framework for event-based frame interpolation with ad-hoc deblurring in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [62] Lei Sun, Andrea Alfarano, Peiqi Duan, Shaolin Su, Kaiwei Wang, Boxin Shi, Radu Timofte, Danda Pani Paudel, Luc Van Gool, Qinglin Liu, et al. Ntire 2025 challenge on event-based image deblurring: Methods and results. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1324–1341, 2025. 2
- [63] Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. Deep learning in spiking neural networks. *Neural networks*, 111: 47–63, 2019. 1
- [64] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proc. CVPR*, pages 16155–16164, 2021. 2
- [65] 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. CVPR*, pages 17755–17764, 2022. 2, 6, 7
- [66] Patricia Vitoria, Stamatios Georgoulis, Stepan Tulyakov, Alfredo Bochicchio, Julius Erbach, and Yuanyou Li. Event-based image deblurring with dynamic motion awareness. *arXiv preprint arXiv:2208.11398*, 2022. 2
- [67] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3
- [68] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proc. CVPR*, pages 10081–10090, 2019. 2
- [69] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025. 3
- [70] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2543–2552, 2021. 6, 7, 8
- [71] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37:92529–92553, 2024. 3, 4
- [72] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25456–25467, 2024. 7
- [73] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proc. ICCV*, 2021. 2
- [74] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4): 1–39, 2023. 3
- [75] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4): 1–39, 2023. 2
- [76] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024. 3
- [77] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024. 3
- [78] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild, 2024. 4
- [79] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2, 4
- [80] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proc. CVPR*, pages 17765–17774, 2022. 2

- [81] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. [2](#), [5](#), [6](#)
- [82] Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multi-vehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. [7](#)
- [83] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018. [2](#), [4](#)