Noname manuscript No.
(will be inserted by the editor)

E2VIDiff: Perceptual Events-to-Video Reconstruction using Diffusion Priors

Jinxiu Liang†, Bohan Yu†, Yixin Yang, Yiming Han, and Boxin Shi ✉

Received: date / Accepted: date

Abstract Event cameras, mimicking the human retina, capture brightness changes with unparalleled temporal resolution and dynamic range. Integrating events into intensities poses a highly ill-posed challenge, marred by initial condition ambiguities. Traditional regression-based deep learning methods fall short in perceptual quality, offering deterministic and often unrealistic reconstructions. In this paper, we introduce diffusion models to events-to-video reconstruction, achieving colorful, realistic, and perceptually superior video generation from achromatic events. Powered by the image generation ability and knowledge of pretrained diffusion models, the proposed method can achieve a better trade-off between the perception and distortion of the reconstructed frame compared to previous solutions. Extensive experiments on benchmark datasets demonstrate that our approach can produce diverse, realistic frames with faithfulness to the given events.

Keywords Events-to-video reconstruction · Diffusion models · Generative modeling · Event camera

1 Introduction

Known as the 'silicon retina', event cameras mimic the human perception system's response to asynchronous, per-pixel logarithmic brightness changes rather than intensity levels periodically. They enjoy unparalleled advantages over conventional frame-based cameras in sensing fast motion of the order of microseconds and a dynamic range larger than 120 dB (Lichtsteiner et al, 2008), which benefit various machine vision applications (Mitrokhin et al, 2018; Vidal et al, 2018). Despite their advantages, signals from event cameras have an appearance less intuitive for human perception, which restricts its usage in contexts demanding human-computer collaboration and sophisticated visual analysis—tasks where subjective visual experience is also required and currently beyond the sole capability of machines. This limitation also hinders the integration of advancements in frame-based machine vision, propelled by deep learning and extensive visual data annotations, with event-based machine vision.

Encapsulating the visual signal in a highly compressed form, events hold the potential for 'events-to-video' reconstruction with arbitrarily high frame rates and dynamic range (Bardow et al, 2016) to help bridge event cameras to human perception and frame-based machine vision. However, it is intrinsically an ill-posed problem that can have an infinite number of valid solutions for given events. Integrating per-pixel brightness changes during a time interval results in only increment images. To obtain the absolute brightness at the end of the interval, an offset image recording the brightness at the beginning of the interval is also needed (Gallego et al, 2022). In addition, the noise model of commercial event cameras deviates from that of conventional cameras. Color information is pivotal in human perception, yet the majority of commercially available event cameras primarily capture achromatic (black and white) events.

Early efforts in converting events to video relied on hand-crafted priors within constrained scene and motion contexts (Kim et al, 2014; Bardow et al, 2016; Munda et al, 2018; Scheerlinck et al, 2019b). Recent approaches have shifted towards deep learning paradigms, utilizing ground

† Equal contribution.

✉ Boxin Shi (Corresponding author)
E-mail: shiboxin@pku.edu.cn

Jinxiu Liang · Bohan Yu · Yixin Yang · Boxin Shi
National Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, China.

Yiming Han
School of Software and Microelectronics, Peking University, China.

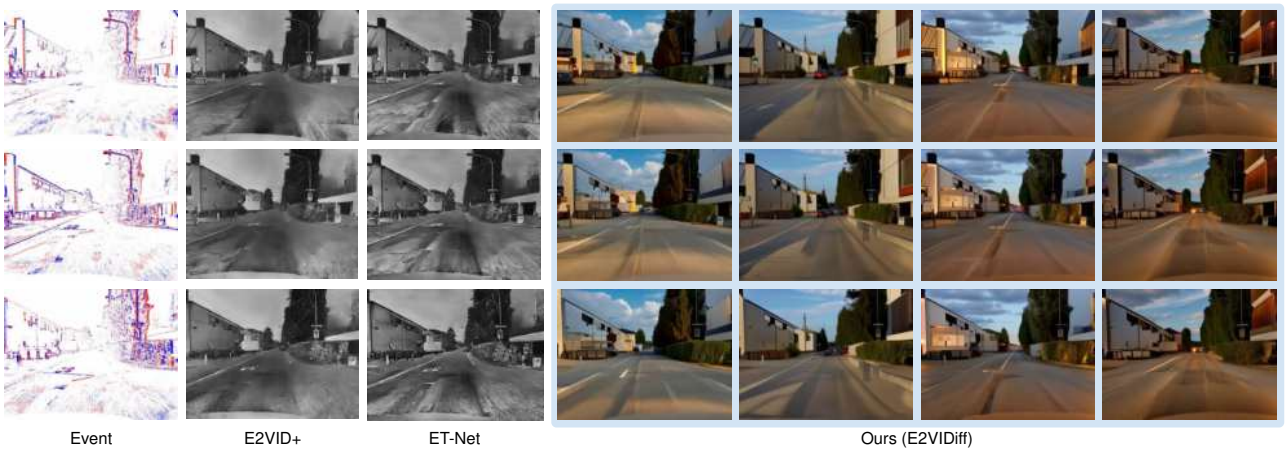


Fig. 1 Visual comparison results of state-of-the-art events-to-video methods E2VID+ (Stoffregen et al, 2020), ETNet (Weng et al, 2021) and the proposed E2VIDiff. Given achromatic events, our approach generates high-quality chromatic videos that are perceptually accurate and aligned with the input events, offering photorealism, vivid coloration, and diversity. Please refer to the supplementary *video* for video reconstruction comparisons.

truth supervision to decrease reliance on restricted assumptions and more effectively capture the intricacies of the real world (Rebecq et al, 2019; Scheerlinck et al, 2020; Stoffregen et al, 2020; Paredes-Valles and de Croon, 2021; Weng et al, 2021; Rebecq et al, 2021; Zhu et al, 2022). However, these regression-based methods often yield deterministic output, a fundamental mismatch for the inherently ill-posed events-to-video reconstruction challenge. They struggle to handle the uncertainty and variability inherent in the unknown initial conditions, noisy event signal, and absent color information.

In situations where several feasible latent frames correspond to the same input events, conventional loss functions often lead reconstructions to converge towards an average of these outcomes, suffering from the *regression to the mean* issue. For example, reconstructing frames of cars driving against diverse street backgrounds from input events can yield multiple valid reconstructions that might reflect cars against streets with varying textures and colors, possibly drawn from Gaussian distributions. Yet, the reconstruction that minimizes the mean squared error typically results in a less detailed, faded, blurred, and low-contrast background - essentially, a conditional mean of latent frames given the events (refer to the roadway and the surrounding vegetation in the results of compared methods in Fig. 1).

To address the issue, a promising strategy involves training the model to sample various plausible reconstructions given the events from the posterior distribution, instead of performing point estimations by finding its maximum or calculating its mean. Recently, diffusion models trained on large-scale dataset become powerful prior models of natural images (Ho et al, 2020; Rombach et al, 2022) and videos (Guo et al, 2023; Luo et al, 2023), which explicitly characterize the conditional data distribution and allow posterior sampling in the image and video manifold. Despite their unique properties, utilizing pretrained diffusion models to address the aforementioned regression to the mean issue in events-to-video reconstruction is nontrivial, due to the following challenges. First, models trained on large-scale datasets of images and videos have a large *modality gap* with events, which requires massive training with prohibitive expense for effective generalization. However, the current data amount of events is much scarcer than that of images and videos. Second, generative diffusion models risk *compromising faithfulness* to original events, exacerbated by temporal inconsistencies and artifacts inherent in event signals stemming from quantization effects and deviations in the noise model.

In this work, we approach events-to-video reconstruction as a conditional generative modeling problem, introducing **E**vent-**To**-**Vi**deo **Diff**usion models called **E2VIDiff** trained to reconstruct latent frames from input events, guided by known ground truth. Our approaches to overcome challenges in the generative events-to-video process include: *i)* employing off-the-shelf video diffusion models pretrained on large-scale dataset, reducing training data requirements while ensuring temporal consistency; *ii)* developing a spatiotemporally factorized event encoder and fusion modules for multimodal and temporal integration, which effectively bridge the modality gap; *iii)* introducing an event-guided sampling mechanism that directs the denoising process for higher-fidelity reconstructions. Experiments demonstrate that this method achieves colorful, photorealistic, and faithful reconstructions from input achromatic events (refer to the results of ours in Fig. 1), offering a novel solution for events-to-video reconstruction. To summarize, our contributions are as follows:

- pioneering diffusion models for color video reconstruction from achromatic events;
- efficient modules for bridging the gap between pretrained diffusion models and events; and
- an event-guided sampling mechanism for faithful events-to-video reconstructions.

2 Related Work

Event-to-video reconstruction. With the unique advantage of high temporal resolution and high dynamic range, event cameras benefit various vision applications such as tracking (Gehrig et al, 2020), deblurring (Zhang et al, 2022; Zhou et al, 2023), and depth estimation (Rebecq et al, 2018; Mostafavi et al, 2021). In recent years, there has been a significant focus on image reconstruction from events. Early reconstruction approaches relied on hand-crafted priors to estimate the intensities from given events, employing techniques such as optimization (Bardow et al, 2016), regularization (Munda et al, 2018), and temporal filtering (Scheerlinck et al, 2019b,a). More recently, deep learning methods have emerged as powerful alternatives for events-to-video reconstruction. These methods alleviate the need for specific assumptions about the scene or motion, enabling the handling of more complex scenarios. By utilizing recurrent U-Net architectures (Rebecq et al, 2019, 2021; Cadena et al, 2021; Zou et al, 2021), generative adversarial networks (GANs) (Wang et al, 2019), fully convolutional networks with recurrent layers (Scheerlinck et al, 2020), transformer models (Weng et al, 2021), and spiking neural networks (Zhu et al, 2022), approaches performing point estimation have demonstrated impressive performance in this regard. To address the limitation of training data diversity, the use of advanced synthetic datasets has been proposed (Stoffregen et al, 2020), enhancing the generalization ability of these networks in varied scenes and motions. These deep learning approaches mark a significant advancement over traditional methods, offering more robust and adaptable solutions for events-to-video reconstruction. However, they tend to produce results with reduced detail, attributed to the regression to the mean issue inherent in point estimators.

Conditional diffusion models. Recently, diffusion models have gained significant attention in image modeling due to their robust training objectives and scalability. Conditional generation with diffusion models has made notable progress, with approaches that leverage paired data and conditioning information during training. For example, in Dhariwal and Nichol (2021), a classifier is trained in parallel with the denoiser to guide the model in generating images with specific attributes. Similarly, Choi et al (2021) conditions the generation on an auxiliary image to guide the denoising process. There have been works such as Song et al (2022); Kawar et al (2021); Wang et al (2022) that utilized pretrained unconditional diffusion models as priors to solve linear inverse imaging problems. These approaches modified the diffusion model sampling algorithm with knowledge of the linear degradation operator to reconstruct images consistent with the learned prior and the given measurements. Text-to-image synthesis is another active field in diffusion modeling. Works such as Ramesh et al (2022); Nichol et al (2022) have pushed the quality of synthesized images to new levels. Latent diffusion models (Rombach et al, 2022) apply diffusion models on lower resolution encoded signals, showing competitive quality with improved speed and efficiency. Imagen (Saharia et al, 2022) achieves remarkable performance in text-to-image synthesis by using large language models for text processing. Versatile diffusion (Xu et al, 2023) further unifies text-to-image, image-to-text, and their variations within a single multiflow diffusion model. There are fewer works in the domain of video modeling due to the computational cost associated with training on video data and the limited availability of large-scale, publicly available video datasets. Nevertheless, there have been notable contributions, such as Voleti et al (2022); Khachatryan et al (2023); Wu et al (2023); Blattmann et al (2023); Guo et al (2023); Luo et al (2023), which address various aspects of video modeling and synthesis. Currently, there are no existing methods that consider the problem of reconstructing color videos from achromatic events.

3 Method

In this section, we begin by delineating the problem definition (Sec. 3.1) of events-to-video reconstruction. Subsequently, an overview of the proposed events-to-video diffusion models E2VIDiff for posterior sampling is presented (Sec. 3.2), which precedes a detailed exposition on the reverse sampling strategy to improve faithfulness to given events (Sec. 3.3) and the network architecture tailored to efficiently bridge the modality gap (Sec. 3.4).

3.1 Problem definition

An event $e_t = (t, j, \sigma)$ at the pixel $j = (j_x, j_y)^\top$ and time t is triggered whenever the logarithmic change of irradiance L exceeds a predefined threshold θ (> 0),

$$\|\log L_t^{(j)} - \log L_{t-\delta t}^{(j)}\| \geq \theta, \tag{1}$$

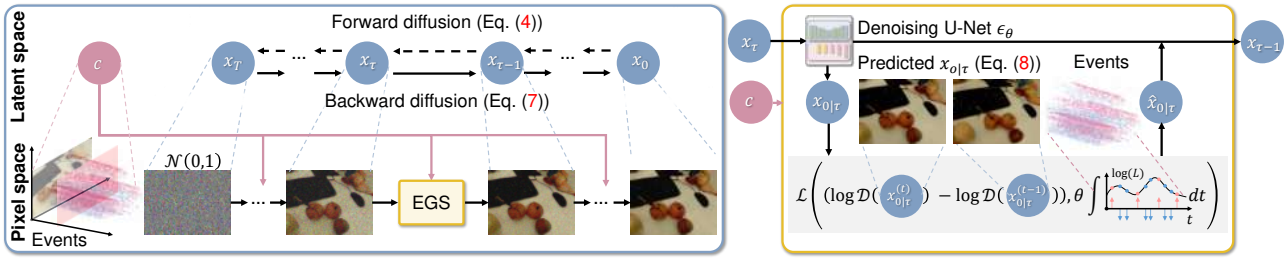
where L_t denotes the instantaneous intensity at time t, and the polarity $\sigma \in \{-1, +1\}$ indicates $\{$negative, positive$\}$ brightness changes. Events can be integrated to reconstruct the absolute brightness image of the scene as

$$\log L_t = \log L_{t_{\text{offset}}} + \theta E_{t_{\text{offset}} \to t}, \tag{2}$$

where $\log L_t$ denotes the logarithmic latent frame, and

$$E_{t_{\text{offset}} \to t} = \int_{t_{\text{offset}}}^{t} e_\tau d\tau. \tag{3}$$

It can be seen that integration of per-pixel brightness changes during a time interval produces only the brightness *increment image* $E_{t_{\text{offset}} \to t}$, and an *offset image* $L_{t_{\text{offset}}}$ (*i.e.*, the brightness image at the start of the interval) needs to be added to



(a) Unfolded diffusion process of the proposed events-to-video diffusion models (b) Event-Guided Sampling (EGS)

Fig. 2 Schematic of the proposed E2VIDiff. (a) The process begins with the latent frame representation x_0 derived from frame L, which undergoes a forward diffusion via the transition kernel expressed in Eq. (4). Conversely, backward diffusion begins from pure Gaussian noise $x_T \sim \mathcal{N}(0, I)$ and iteratively reconstructs a predicted sample $x_0 \sim p(x|c)$ by a backward diffusion expressed in Eqs. (7) and (8). (b) The proposed event-guided sampling mechanism iteratively enhances the fidelity and perceptual quality of the generated frame by alternating between denoising (to align with natural image characteristics) and ensuring the predicted sample $x_{0|\tau}$ is consistent with the physical formulation of the given events.

obtain the absolute brightness L_t at the end of the interval. Moreover, suppose $\widetilde{L}_t$ is a valid solution of frame at timestamp t for certain offset image $\widetilde{L}_{t_{\text{offset}}}$, then $\alpha\widetilde{L}_t$ is also an valid solution corresponds to offset image $\alpha\widetilde{L}_{t_{\text{offset}}}$ for any nonzero α, including the case where α is spatially varying. Therefore, solving the events-to-video problem often requires regularization techniques to incorporate prior knowledge of the appearance of a natural image. Priors are also necessary to account for quantization effects and noise model deviated from conventional cameras in real-world scenarios, and to prevent overfitting to specific scenes or lighting conditions.

Consider samples $\mathbb{L}, \mathbb{E}$ drawn from an unknown distribution $p(\mathbb{L}, \mathbb{E})$, where $\mathbb{L} = \{L^{(t)}\}_{t=0}^{N-1}$ denotes a sequence of N sharp chromatic frames, and $\mathbb{E} = \{E^{(t)}\}_{t=0}^{N-1}$ denotes a stack of event stream partitioned into N sequential spatiotemporal windows corresponding to the exposure time of the N frames. The goal of events-to-video reconstruction is to reconstruct a sequence of frames $\mathbb{L}$ from their corresponding event stream $\mathbb{E}$. In our work, it is treated as a conditional generative problem, where the conditional distribution $p(\mathbb{L}|\mathbb{E})$ is a one-to-many mapping in which many target images may be consistent with given events. We are interested in learning a parametric approximation to $p(\mathbb{L}|\mathbb{E})$ through a stochastic translation process that maps given event stream $\mathbb{E}$ to target frame sequence $\mathbb{L}$ using diffusion models.

3.2 Events-to-video diffusion models

Our approach leverages diffusion models for posterior sampling in the video manifold, facilitating the reconstruction of diverse, colorful, and perceptually high-quality videos from achromatic input events. The overall pipeline of the proposed E2VIDiff is shown in Fig. 2.

Diffusion models perform data modeling via iterative diffusion and reversely denoising, where the denoising network is trained with denoising score matching (Ho et al, 2020). To alleviate the computational burden in the pixel space, the proposed E2VIDiff is based on latent diffusion models (Rombach et al, 2022) that operate the diffusion process in the latent space. Specifically, a compression model is used to transform the frame sequence $\mathbb{L} \sim p_{\text{data}}$ into a lower-dimensional latent space. A variational auto-encoder that maximized the lower bound of the evidence (Esser et al, 2021) leads to the latent presentation $x = \mathcal{E}(\mathbb{L})$, where $\mathcal{E}$ is the encoder. The resulting pixel domain representation can be obtained by $\widehat{\mathbb{L}} = \mathcal{D}(x)$, where $\mathcal{D}$ is the decoder.

During the training phase, the target frame sequence $\mathbb{L}$ is initially mapped to the latent space via the frozen encoder $\mathcal{E}$, resulting in $x_0 = \mathcal{E}(\mathbb{L})$. The forward process begins with the representation x_0 that are repeatedly diffused by the transition kernel $q(x_\tau|x_{\tau-1})$ as follows:

$$q(x_\tau|x_{\tau-1}) = \mathcal{N}(x_\tau; \sqrt{\alpha_\tau}x_{\tau-1}, (1-\alpha_\tau)I), \tag{4}$$

where I denotes identity matrix, until it reaches the pure Gaussian noise $x_T \sim \mathcal{N}(0, I)$ given a sufficiently large step T. Here, $0 < \alpha_\tau < 1$ comprises a noise schedule, parameterized by a diffusion step τ, which causes the logarithmic signal-to-noise ratio $\lambda_\tau = \log(\alpha_\tau/(1-\alpha_\tau))$ to decrease monotonically. It results in closed-form expressions for a partially noisy latent x_τ at an arbitrary step τ as:

$$x_\tau = \sqrt{\gamma_\tau}x_0 + \sqrt{1-\gamma_\tau}\epsilon_\tau, \quad \epsilon_\tau \sim \mathcal{N}(0, I), \tag{5}$$

where $\gamma_\tau = \prod_{j=1}^{\tau} \alpha_j$, making the training of a diffusion probabilistic model practical. Accordingly, a conditional denoising model $\epsilon_\theta(x_\tau; c, \tau)$ parameterized by parameters θ is trained to estimate ϵ from the diffused x_τ with the conditioning information c extracted from the given event stream $\mathbb{E}$ as a condition. The denoising is optimized by minimizing the following ℓ_2 noise prediction objective:

$$\mathbb{E}_{x_0 \sim \mathcal{E}(\mathbb{L}), \tau \sim p_\tau, \epsilon_\tau \sim \mathcal{N}(0,I)}[\|\epsilon_\tau - \epsilon_\theta(x_\tau; c, \tau)\|_2^2], \tag{6}$$

where p_τ are a uniform distribution over the diffusion time τ. The detailed network architecture of $\epsilon_\theta(x_\tau; c, \tau)$ is introduced in Sec. 3.4.

Upon completion of the training, the reverse process runs backward from pure Gaussian noise x_T to a predicted sample $x_0 \sim p(x|c)$. The maximum diffusion time is usually selected so that the input data are completely diffused into Gaussian random noise, which we employ $p_\tau \sim \mathcal{U}\{0, 1000\}$ in the method. Assuming deterministic DDIM sampling as described by Song et al (2020), reverse sampling from $p_\theta(x_{\tau-1}|x_\tau)$ is conducted as

$$x_{\tau-1} = \sqrt{\gamma_{\tau-1}}\widetilde{x}_{0|\tau} + \sqrt{1-\gamma_{\tau-1}}\epsilon_\theta(x_\tau; c, \tau), \tag{7}$$

where

$$\widetilde{x}_{0|\tau} = \left(x_\tau - \sqrt{1-\gamma_\tau}\epsilon_\theta(x_\tau; c, \tau)\right)/\sqrt{\gamma_\tau}. \tag{8}$$

Denoised estimates of the desired frame sequence on pixel space can be obtained by

$$\widetilde{\mathbb{L}}_\tau = \mathcal{D}(\widetilde{x}_{0|\tau}). \tag{9}$$

To improve faithfulness to given events, we impose the consistency between the predicted consecutive frames $\mathbb{L}$ and the observed events $\mathbb{E}$ in the event-guided sampling mechanism detailed in Sec. 3.3.

3.3 Event-guided sampling

Generative diffusion models may compromise the faithfulness to given events, often due to temporal discrepancies and artifacts from quantization and deviations in the noise model. Although capable of creating realistic frames, they may produce outputs that deviate significantly from the given events. To mitigate this, we introduce the event-guided sampling (EGS) mechanism, designed to steer the denoising process towards high-fidelity video reconstructions. Specifically, by using the consistency between the predicted frames $\mathbb{L}$ and the given events $\mathbb{E}$ as guidance, EGS iteratively balances the distortion and perceptual quality of the generated frames and progressively projects the samples into the video manifold. The process of EGS is shown in the right of Fig. 2.

Consider the event stack $E^{(t)}$ integrated via Eq. (3) from the event stream $\{e_k = (t_k, j_k, \sigma_k)\}_{k=0}^{K}$ with threshold θ triggered during the exposure time between adjacent frames with indexes $t-1$ and t. Denote the distance between $E^{(t)}$ and the brightness difference between consecutive frames $\widetilde{L}_\tau^{(\cdot)}$ obtained from $x_\tau^{(\cdot)}$ by Eqs. (8) and (9) as:

$$d_\tau^{(t)} = \log \widetilde{L}_\tau^{(t)} - \log \widetilde{L}_\tau^{(t-1)} - \theta E^{(t)}. \tag{10}$$

According to Eq. (1), a feasible solution satisfies:

$$-2\theta \le d_\tau^{(t)} \le 2\theta, \tag{11}$$

which can be reformulated into the following loss function where the gradient with respect to $x_\tau^{(t)}$ can be computed:

$$\mathcal{L}_{\text{event}} = \max(0, \|d_\tau^{(t)}\|_1 - 2\theta). \tag{12}$$

Algorithm 1 Event-guided sampling

Required: Encoder $\mathcal{E}$, decoder $\mathcal{D}$, U-Net ϵ_θ, event encoder $\mathcal{E}_{\text{E}}$
Input: Maximum diffusion time T, number of optimization step S, number of frame in a sequence N, event stack sequence $\{E^{(t)}\}_{t=1}^{N}$, noise schedule $\{\gamma_\tau\}_{\tau=0}^{T}$
Output: $\{\mathcal{D}(x_0^{(t)})\}_{t=0}^{N}$

for $t = 0$ to N **do**
  $x_T^{(t)} \sim \mathcal{N}(0, I)$
  $c^{(t)} = \mathcal{E}(E^{(t)})$
end for
for $\tau = T$ to 1 **do**
  for $t = 0$ to N **do**
    $\widehat{x}_\tau^{(t)} \leftarrow x_\tau^{(t)}$
    $c = \mathcal{E}_{\text{E}}(E^{(t)})$
    $\widehat{\epsilon}_\tau^{(t)} \leftarrow \epsilon_\theta(x_\tau^{(t)}; c^{(t)}, \tau)$
    for $s = 0$ to $S-1$ **do**
      $\widetilde{x}_{0|\tau}^{(t)} \leftarrow \left(x_\tau^{(t)} - \sqrt{1-\gamma_\tau}\widehat{\epsilon}_\tau^{(t)}\right)/\sqrt{\gamma_\tau}$ ▷ Eq. (8)
      $\widetilde{L}_\tau^{(t)} \leftarrow \mathcal{D}(\widetilde{x}_{0|\tau}^{(t)})$ ▷ Eq. (9)
      if $t \ge 1$ **then**
        $d_\tau^{(t)} \leftarrow \log \widetilde{L}_\tau^{(t)} - \log \widetilde{L}_\tau^{(t-1)} - \theta E^{(t)}$ ▷ Eq. (10)
        $\mathcal{L} \leftarrow \max\left(0, \|d_\tau^{(t)}\|_1 - 2\theta\right) + \eta\|x_\tau^{(t)} - \widehat{x}_\tau^{(t)}\|_2^2$ ▷ Eqs. (12), (13) and (14)
        $x_\tau^{(t)} \leftarrow \text{LBFGS}(\mathcal{L}, x_\tau^{(t)})$
      end if
    end for
    $\widetilde{x}_{0|\tau}^{(t)} \leftarrow \left(x_\tau^{(t)} - \sqrt{1-\gamma_\tau}\widehat{\epsilon}_\tau^{(t)}\right)/\sqrt{\gamma_\tau}$ ▷ Eq. (8)
    $x_{\tau-1}^{(t)} \leftarrow \sqrt{\gamma_{\tau-1}}\widetilde{x}_{0|\tau}^{(t)} + \sqrt{1-\gamma_{\tau-1}}\widehat{\epsilon}_\tau^{(t)}$ ▷ Eq. (7)
  end for
end for

This loss function imposes an ℓ_1 regularizations on the logarithmic brightness difference between consecutive frames $\mathbb{L}$ indexed by $t = 0, 1, ..., N$ to push the solution into the feasible region that consistent to the observation of events $\mathbb{E}$.

Moreover, to avoid that the solution getting too far away from the feasible space of diffusion model, an ℓ_2 loss between the guided sampling prediction $x_\tau^{(t)}$ and the solution $\widehat{x}_\tau^{(t)}$ sampled from the last DDIM step:

$$\mathcal{L}_{\text{diff}} = \|x_\tau^{(t)} - \widehat{x}_\tau^{(t)}\|_2^2. \tag{13}$$

At each timestamp τ, the overall constraints are formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{event}} + \eta\mathcal{L}_{\text{diff}}, \tag{14}$$

where η is to balance the weight of different terms, and $x_\tau^{(t)}$ is updated under the guidance of $\mathcal{L}$ by the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) (Liu and Nocedal, 1989) optimizer for S steps. Under this EGS mechanism, the reverse process Eq. (7) is performed iteratively to progressively remove noise and obtain denoised results that are realistic and faithful to the given events. Algorithm 1 outlines such EGS mechanism.

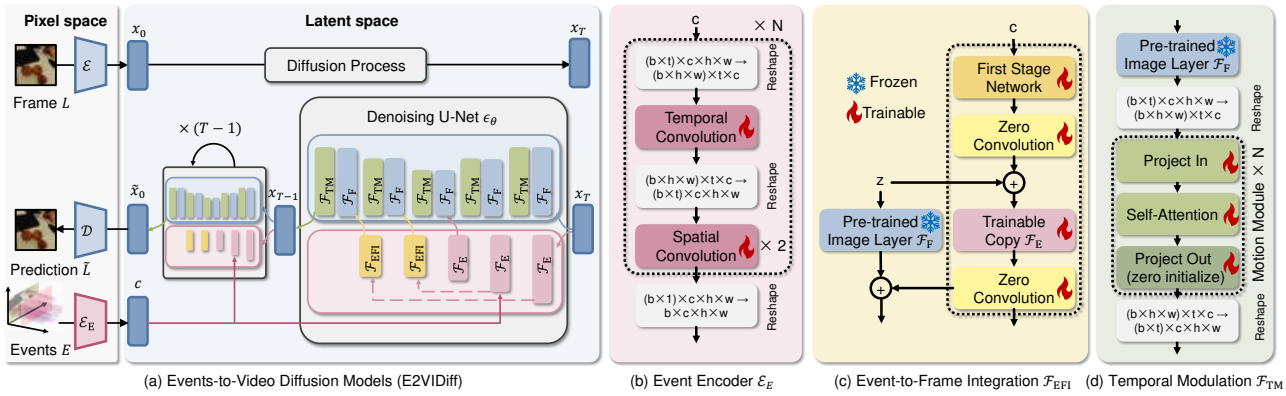


Fig. 3 The network architecture of the denoising U-Net ϵ_θ in the proposed E2VIDiff. (a) Overview of the denoising process utilizing the denoising U-Net ϵ_θ, featuring dual branches with weights initialized from pretrained image diffusion models. (b) The event extractor $\mathcal{E}_\text{E}$ is designed to extract conditioning information from the event stream. (c) The event-to-frame integration module $\mathcal{F}_\text{EFI}$ is designed to merge event features with the output of the frozen image diffusion models. (d) The temporal modulation module $\mathcal{F}_\text{TM}$ ensures temporal consistency in the generated frames.

3.4 Network architecture

To mitigate the modality gap of pretrained diffusion models with events, and the limited availability of event data, we develop a spatiotemporally factorized event encoder (EE). Together with an event-to-frame integration (EFI) module, the proposed method effectively bridges the gap between different data modalities. Temporal modulation (TM) module is employed to efficiently enhance temporal consistency. The overall network architecture designs are shown in Fig. 3.

The denoising U-Net ϵ_θ consists of two branches, both with U-Net-like architectures and weights initialized from pretrained image diffusion models as shown in Fig. 3(a). The frame branch $\mathcal{F}_\text{F}(x_\tau;\tau)$ accounts for sampling the features of plausible offset image (corresponding to L_{t_offset} in Eq. (2)) from the random latent x_T sampled from normal space, while the event branch $\mathcal{F}_\text{E}(x_\tau;c,\tau)$ accounts for generating the features of increment image (corresponding to $E_{t_\text{offset}\to t}$ in Eq. (2)) from the conditioning information c extracted from the input events by the EE module $\mathcal{E}_\text{E}$.

To align with the encoder $\mathcal{E}$ used by the pretrained image diffusion models, the EE module $\mathcal{E}_\text{E}$ convert the event $E^{(t)}$ into a 64×64 feature $c = \mathcal{E}_\text{E}(E^{(t)})$ for an input of size 512×512 to match the size of the convolutional layer as shown in Fig. 3(b), where c represents conditioning information. It contains six space-time-factorized event encoder blocks. Each event encoder block consists of a 1D temporal convolution following each 2D spatial convolution layer. This enables efficient sharing of information across spatial and temporal dimensions, avoiding the intensive computational demands of 3D convolutional layers. Moreover, it distinctly separates the established 2D convolutional layers that can utilize off-the-shelf encoder for similar modalities from the newly introduced 1D convolutional layers. This separation permits the training of the temporal convolutions anew, while preserving the spatial understanding previously acquired in the pretrained weights of the spatial convolutions. For effective initialization, the temporal convolution layer starts as an identity function. This strategy ensures a smooth transition from pretrained layers focused solely on spatial aspects to those handling both spatial and temporal elements.

To efficiently integrate the conditioning information from events to frames, the information from the frame branch $\mathcal{F}_\text{F}$ and the event branch $\mathcal{F}_\text{E}(x_\tau;c,\tau)$ are fused by the EFI module $\mathcal{F}_\text{EFI}$ as shown in Fig. 3(c). It aligns the feature of events in the low-dimensional latent space defined by pretrained latent diffusion models for natural frames, bridging the gap between different data modalities. The TM module $\mathcal{F}_\text{TM}$ is employed for temporal consistency in frame generation, requiring 5D tensors for input and undergoing a model inflation process for adaptation from pretrained image diffusion models. As shown in Fig. 3(d), it transforms 2D layers $\mathcal{F}_\text{F}$ into pseudo-3D, utilizing self-attention blocks for temporal dependency capture, with zero-initialized output layers for smooth integration. This mechanism ensures motion smoothness and content consistency across generated frames.

4 Experiments

4.1 Experimental setup

Dataset[1]. The quantitative evaluation is carried out on three public datasets: HQF (Stoffregen et al, 2020), IJRR (Mueggler et al, 2017), MVSEC (Zhu et al, 2018). The HQF (Stoffregen et al, 2020) dataset comprises 14 sequences captured by the DAVIS240 camera. The IJRR (Mueggler et al, 2017) dataset includes 25 realistic sequences captured with the DAVIS240 camera. In contrast, the MVSEC (Zhu et al, 2018) dataset is recorded using a synchronized stereo event camera

[1] All datasets used for evaluation and training are publicly accessible.

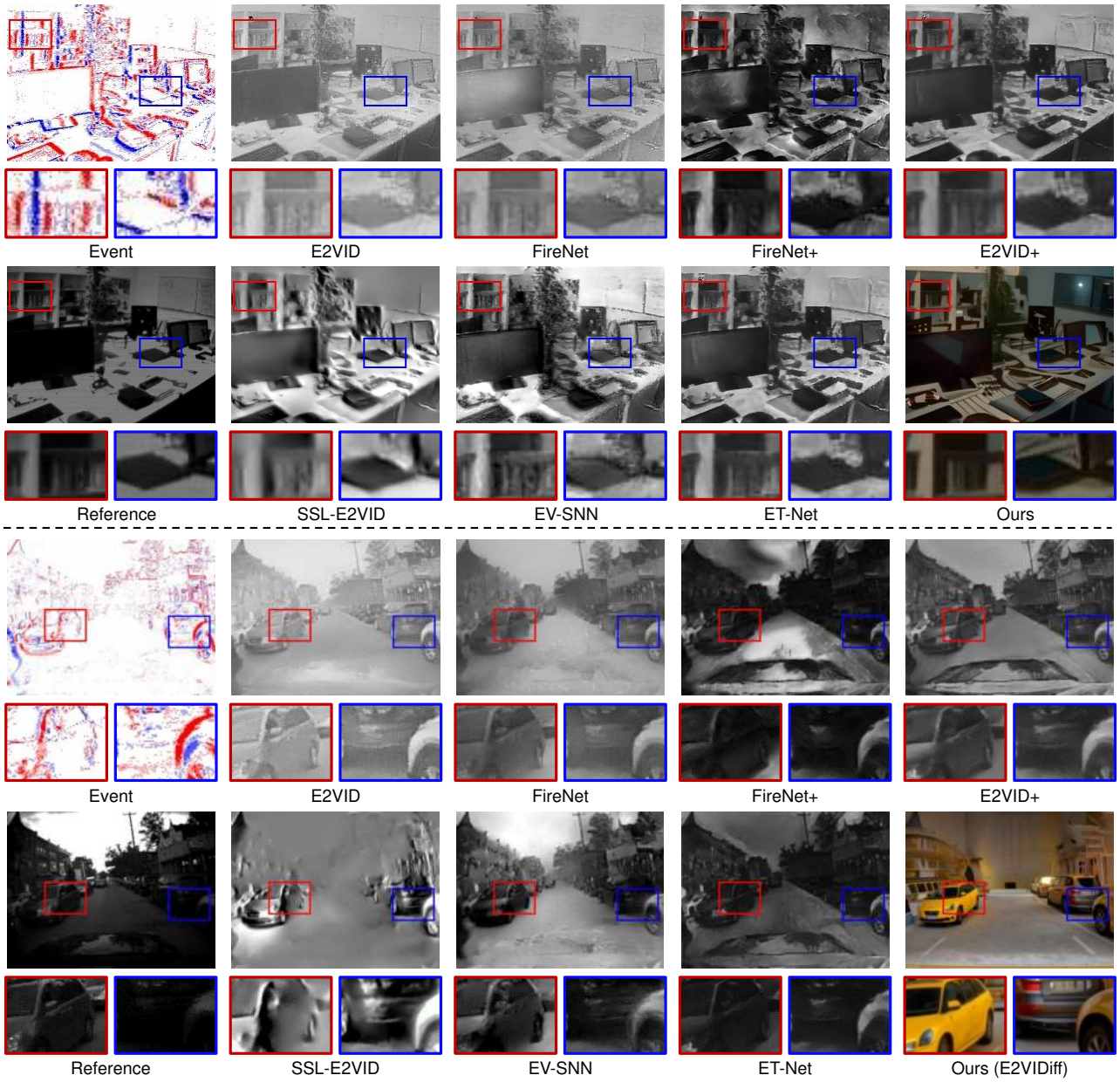


Fig. 4 Qualitative comparisons on real events captured by DAVIS240 (Mueggler et al, 2017) (top) and DAVIS346B (Zhu et al, 2018) (bottom) event cameras, with corresponding achromatic frames captured concurrently as references. The compared state-of-the-art methods include: E2VID (Rebecq et al, 2019), FireNet (Scheerlinck et al, 2020), FireNet+ (Stoffregen et al, 2020), E2VID+ (Stoffregen et al, 2020), SSL-E2VID (Paredes-Valles and de Croon, 2021), EV-SNN (Zhu et al, 2022), and ETNet (Weng et al, 2021).

system. All these are real-world event data with the corresponding achromatic frames. For qualitative evaluation and applications to several high-level downstream tasks, we additionally leverage the DSEC dataset (Gehrig et al, 2021a) captured using a synchronized stereo event camera system, in which the events are captured by Prophosee EVK3. For training, we use a text-video pair dataset WebVid-2.5M (Bain et al, 2021). The video clips in the dataset are sampled at the stride of 5, then resized and center-cropped to the resolution of 512×512. The event streams corresponding to the video clips are generated from the ideal event generation model (1) first, and augmented with hot pixels and missing event to account for the non-ideal effect of the event triggering.

Metrics. Following the evaluation protocol defined in previous works on video diffusion models (Blattmann et al, 2023; Luo et al, 2023; Yu et al, 2023), we present the Inception Score (IS) (Salimans et al, 2016) and Frechet Video Distance (FVD) (Unterthiner et al, 2019) scores for various methods, which are pivotal metrics for evaluating the quality and coherence of generated video content in the field of computer

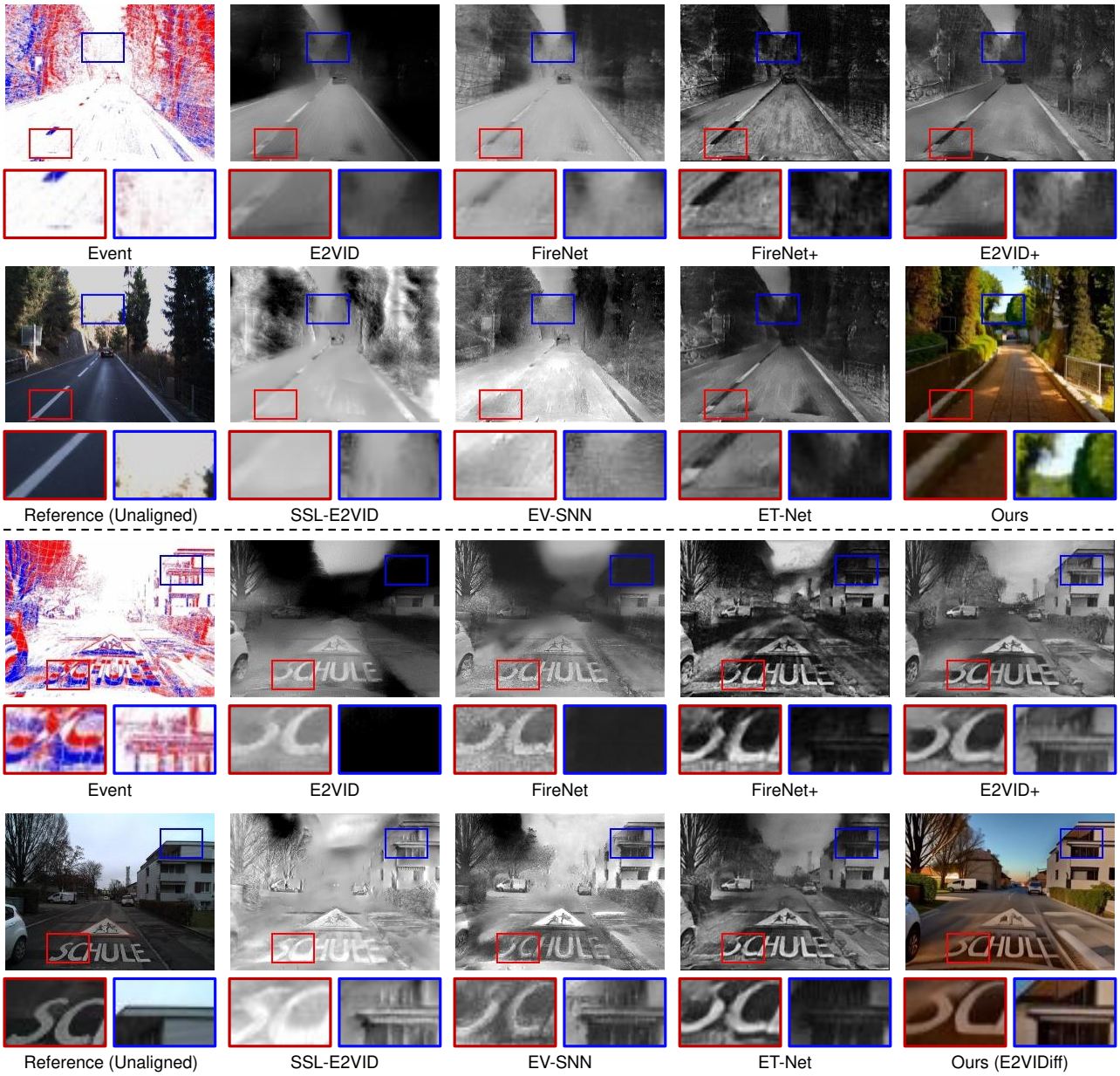


Fig. 5 Qualitative comparisons on real events captured by Prophesee Gen3 (Gehrig et al, 2021a) event cameras, accompanied by chromatic frames captured in stereo alongside the events, serving as unaligned references. Please refer to the caption of Fig. 4 for a complete list of the compared methods.

vision. IS (Salimans et al, 2016) measures the diversity and clarity of the videos generated. A high IS indicates that the model produces varied and distinct videos, each with clearly defined objects or scenes, as perceived by a trained 3D Convolutional Model (C3D) (Tran et al, 2015). The IS is particularly useful for assessing the generative model's ability to produce visually compelling content that aligns with the distribution of real-world videos. FVD (Unterthiner et al, 2019) evaluates the similarity between the distribution of generated videos and real videos. It can capture both the visual and temporal dynamics of videos, making it a comprehensive measure of video quality and authenticity. A lower FVD score signifies a closer resemblance to real video content, as determined by a trained Inflated 3D ConvNet (I3D) model (Carreira and Zisserman, 2017). FVD is crucial for determining the model's effectiveness in capturing the intricacies of video data.

Baselines. For comparison, we compare the proposed method with seven state-of-the-art methods: E2VID (Rebecq et al, 2019), FireNet (Scheerlinck et al, 2020), E2VID+ (Stoffregen et al, 2020), FireNet+ (Stoffregen et al, 2020), SSL-E2VID (Paredes-Valles and de Croon, 2021), EV-SNN (Zhu et al, 2022), and ETNet (Weng et al, 2021). All results are generated using the pretrained models provided in the original

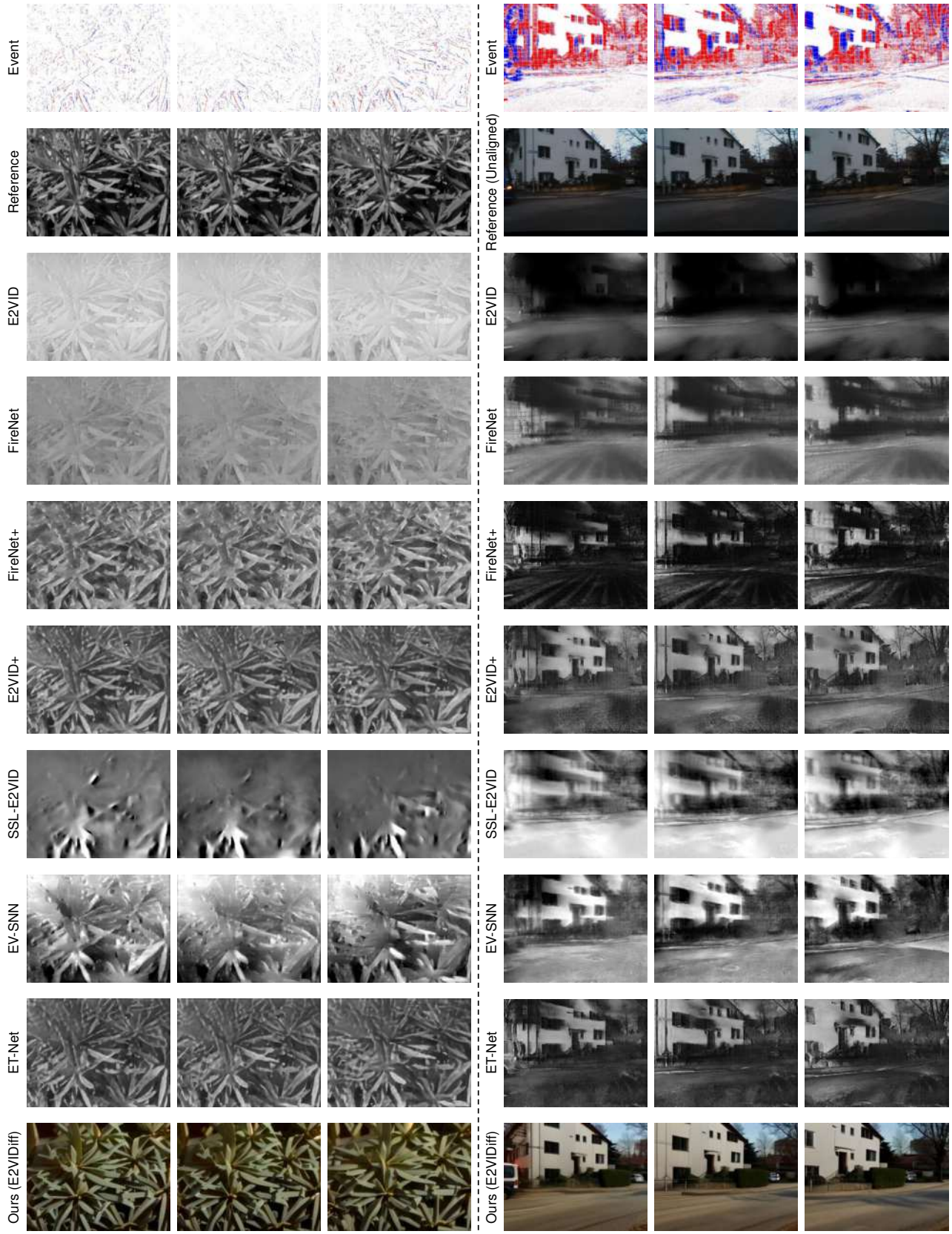


Fig. 6 Qualitative comparisons of temporal consistency in successive frames generated from real events captured by DAVIS240 (Stoffregen et al, 2020) (left) and Prophesee Gen3 (Gehrig et al, 2021a) (right) event cameras, with corresponding achromatic (left) and chromatic (right) frames captured concurrently as references. Please refer to the caption of Fig. 4 for a complete list of the compared methods.

Table 1 Quantitative comparison of IS (↑) between state-of-the-art methods and the proposed E2VIDiff on different datasets with paired real events and frames. Red, orange, and yellow highlights indicate the 1st, 2nd, and 3rd-best performing method. The compared state-of-the-art methods for video reconstruction from events include: E2VID (Rebecq et al, 2019), FireNet (Scheerlinck et al, 2020), FireNet+ (Stoffregen et al, 2020), E2VID+ (Stoffregen et al, 2020), SSL-E2VID (Paredes-Valles and de Croon, 2021), EV-SNN (Zhu et al, 2022), and ETNet (Weng et al, 2021).

Dataset	IJRR	HQF	MVSEC
FireNet	3.62	2.79	2.92
FireNet+	3.55	2.75	3.03
E2VID	3.50	2.61	2.88
E2VID+	3.38	2.78	2.78
SSL-E2VID	3.54	3.14	3.13
EV-SNN	3.45	2.98	2.91
ETNet	3.51	2.95	2.91
Ours (E2VIDiff)	3.79	3.13	3.06

papers. To ensure strict consistency between the timestamps of the reconstruction and the provided ground truth, we employ the events that occur between two adjacent frames to generate each reconstruction.

Implementation details[2]**.** The maximum diffusion time $T = 1000$. The length of the generated frame sequence $N = 16$. The resolution of the video is unrestricted, with a standard resolution of 512×512. We use a 25 step DDIM (Song et al, 2020) sampler for inference and set the classifier-free guidance scale to 7.5. A linear beta schedule is used for γ_τ following Guo et al (2023). The number of optimization steps for EGS is $S = 5$. To maintain strict timestamp consistency between the reconstruction and ground truth, we utilize the events between two adjacent frames to generate each reconstruction. The training of the proposed method is explicitly disentangled into three parts, content, structure, and motion. The frame branch $\mathcal{F}_\text{F}$ for the content is initialized from one of the state-of-the-art image diffusion models (Rombach et al, 2022). The most important part is to ensure that each video frame follows the structure or trajectory of the given events. In the proposed module, it is achieved by the event branch $\mathcal{F}_\text{E}$ while freezing the frame branch $\mathcal{F}_\text{F}$. The weights of $\mathcal{F}_\text{E}$ for per-frame structure-preservation generation are initialized from the HED soft edge variant of Zhang et al (2023). To ensure that the video frames remain temporally coherent, the temporal modulation module $\mathcal{F}_\text{TM}$ is initialized from the pretrained video diffusion model Guo et al (2023). The training process takes place on a single NVIDIA GeForce RTX 4090 24GB GPU and spans a duration of 3 days in total. The learning rate is 1.0×10^{-5} and the batch size is 4.

[2] The source code will be released at https://sherrycattt.github.io/E2VIDiff.

Table 2 Comparison of FVD (↓) between state-of-the-art methods and the proposed E2VIDiff on different datasets with paired real events and frames. Red, orange, and yellow highlights indicate the 1st, 2nd, and 3rd-best performing method. Please refer to the caption of Table 1 for a complete list of the compared methods.

Dataset	IJRR	HQF	MVSEC
FireNet	1373.19	1614.61	2532.41
FireNet+	1438.68	1809.96	2159.91
E2VID	1249.41	2169.55	2622.34
E2VID+	737.21	974.01	2838.63
SSL-E2VID	1799.55	3006.62	2330.01
EV-SNN	1654.94	3177.04	2443.08
ETNet	821.70	1193.30	2618.33
Ours (E2VIDiff)	885.34	768.03	950.52

4.2 Comparisons with state-of-the-arts

Qualitative results. The visual comparison results are shown in Fig. 4 and Fig. 5. Thanks to the ability of posterior sampling of diffusion models, the proposed E2VIDiff achieves the best perceptual quality. This is exemplified by detailed reconstruction of complex scenes, such as the discernible features of bookshelves and the vehicle's head shown in Fig. 4, as well as the tree and house shown in Fig. 5. Furthermore, E2VIDiff demonstrates an exceptional ability to accurately infer and assign plausible colors to various scene elements, as observed in the vibrant coloration of the car at the bottom of Fig. 4. It also excels at capturing scenes characterized by high speed and a wide dynamic range as shown in the driving scenes of Fig. 5, maintaining clarity and detail where other techniques falter. While competing methods struggle to capture the details of the sky, the proposed E2VIDiff retains these elements seamlessly. Comparatively, while E2VID+ (Stoffregen et al, 2020) and ETNet (Weng et al, 2021) emerge as the leading alternatives among the state-of-the-art methods, they exhibit noticeable fog-like artifacts in their results. Furthermore, the inherent limitations of sensor properties and the sparse nature of event steams lead to challenges in generating objects with smooth contours, as evidenced by the jagged contour of the books at the top of Fig. 4.

To demonstrate the quality of temporal consistency, we show successive frame results in Fig. 6. Where other methods exhibit various degrees of temporal discontinuities and jitter caused by unstable artifacts, the proposed E2VIDiff maintains a seamless flow, effectively capturing the motion. This is particularly evident in scenes with rapid movement in Fig. 6, where our method consistently produces stable and coherent reconstructions. The smooth transition between frames, without noticeable artifacts or abrupt changes, highlights the proposed E2VIDiff's proficiency in ensuring temporal consistency, a critical aspect often challenging for video reconstruction tasks. The fidelity of temporal consistency in our reconstructions can be attributed to the synergistic integration of the EFI module and the EGS mechanism, which

work in concert to ensure that each frame not only accurately represents the scene at a given instant, but also coherently aligns with adjacent frames in the temporal domain.

Quantitative results. We use the Inception Score (IS) (Salimans et al, 2016) and Frechet Video Distance (FVD) (Unterthiner et al, 2019) scores as the evaluation metrics following Blattmann et al (2023); Luo et al (2023); Yu et al (2023). In terms of IS (Salimans et al, 2016), SSL-E2VID and the proposed method demonstrates superior performance in terms of IS score, as evidenced in Table 1. The good performance can be attributed to our pioneering use of diffusion models for chromatic video reconstruction from achromatic events. By generating samples from the posterior distribution rather than relying on point estimations, our method introduces a level of diversity and clarity in the generated content that significantly enhances the IS score. For the FVD (Unterthiner et al, 2019) scores, our method achieves state-of-the-art performance across all datasets, with the exception of IJRR, as detailed in Table 2. This outstanding performance underscores the efficacy of our efficient modules designed to bridge the domain gap between pretrained diffusion models and event data. Integration of a spatiotemporally factorized EE module and an EFI module effectively narrows the modality gap, ensuring temporal consistency, and contributing to a lower FVD score. Moreover, our EGS mechanism plays a pivotal role in directing the denoising process, thereby facilitating reconstructions with faithfulness to the given events.

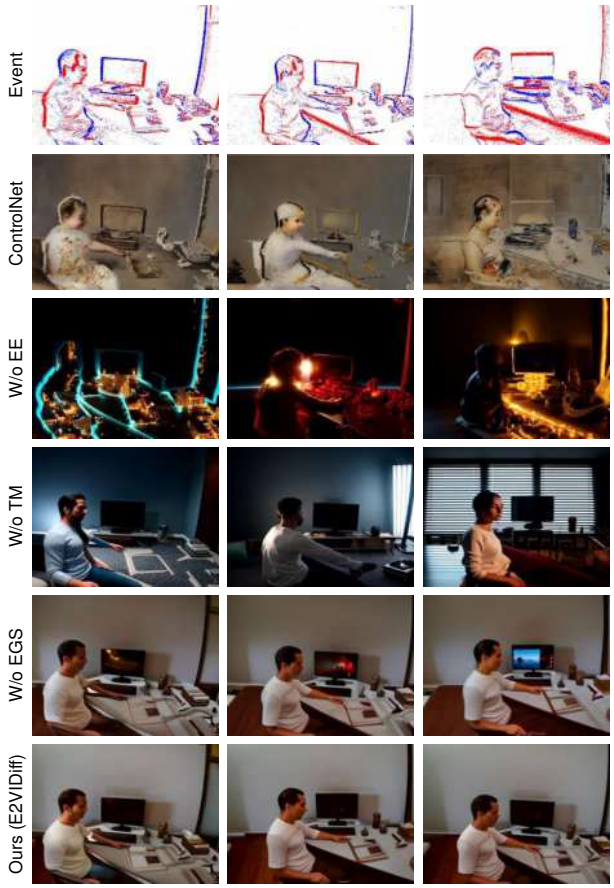


Fig. 7 Ablation study results assessing the individual components of the proposed E2VIDiff. Displayed in sequence from top to bottom are successive frames generated by the following: the baseline ControlNet (Zhang et al, 2023) with Canny edges derived from integrated events (denoted by 'ControlNet'); our method excluding the event encoder (denoted by 'W/o EE'); our method without temporal modulation module (denoted by 'W/o TM'); our method without the event-guided sampling mechanism (denoted by 'W/o EGS'); and the proposed E2VIDiff.

4.3 Ablation studies

In our ablation studies, we evaluate the individual components of our proposed method for chromatic video reconstruction from achromatic events. Qualitative results are illustrated in Fig. 7, showing consecutive frames generated by different configurations of our model.

The second row presents frames generated by the original ControlNet with Canny edges integrated from events as input, which validate the overall contributions of the proposed components of our method. Integrating Canny edges provides a rudimentary form of event data utilization, which falls short of the sophistication and depth of information leveraged by our proposed method.

The comparison between the frames generated by the model without the EE module and those produced by the proposed E2VIDiff highlights the pivotal role of the EE module in capturing the fine details and dynamics present in the raw events. Without the EE module, the model struggles to accurately interpret and utilize the rich information contained within the events, leading to reconstructions that lack detail and exhibit very poor temporal consistency. The EE module effectively translates the sparse and high-dimensional event data into a more manageable form for the diffusion model, enabling a more precise and coherent video reconstruction.

Omitting the TM module results in frames that fail to maintain consistent motion patterns across time, underscoring the module's crucial role in preserving temporal consistency. The TM module integrates motion information derived from consecutive events, facilitating smooth transitions between frames and enhancing the perception of motion in the reconstructed videos. Its absence leads to a disjointed sequence, where the continuity of movement is compromised, detracting from the overall realism of the video.

The impact of the EGS mechanism is evident when comparing its absence with the proposed E2VIDiff. Frames generated without EGS show a noticeable degradation in the fidelity of the reconstructions, with less accurate color representation and diminished clarity. The EGS mechanism directs the denoising process in a manner that faithfully adheres to the underlying event data, ensuring that the generated frames

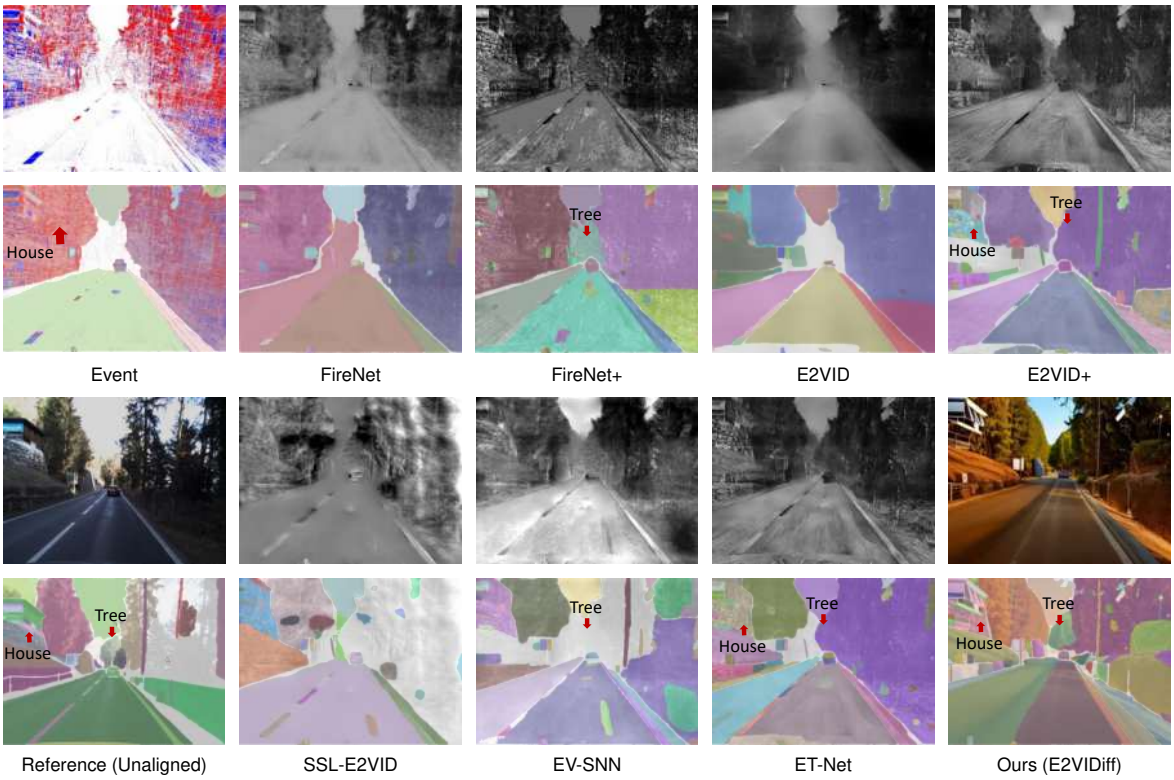


Fig. 8 Semantic segmentation results employing events-to-video reconstructions as an intermediary, enabling standard computer vision algorithms to be applied on event data directly. All methods utilize the SAM (Kirillov et al, 2023) for its notable zero-shot performance, suitable for open-domain semantic segmentation on unaligned reference and frames reconstructed by different methods. Please refer to the caption of Fig. 4 for a complete list of the compared methods.

not only exhibit high visual quality but also accurately reflect the original scene dynamics.

4.4 Results of downstream tasks

We explore using our reconstructions as an intermediary format within the realm of natural images, allowing conventional vision algorithms to be applied directly to events.

Semantic segmentation. We perform semantic segmentation tasks to emphasize the precision and reliability of the proposed E2VIDiff. For all the approaches compared, we employ SAM (Kirillov et al, 2023) for semantic segmentation, which have impressive zero-shot performance that can be applied for open domain. For event data, SAM processes on the integrated event stack. In contrast to reconstructions obtained from Stoffregen et al (2020), our approach yields frames that significantly improve segmentation accuracy. As depicted in Fig. 8, the segmentation of the 'house' region within our reconstructions exhibits a markedly more precise contour, which demonstrates our method's ability to preserve and highlight crucial structural details, thereby enabling segmentation algorithms to perform with greater efficacy.

Optical flow estimation. In the realm of optical flow estimation, we employ E-RAFT (Gehrig et al, 2021b) on event stack, juxtaposed with results generated using RAFT (Teed and Deng, 2020) on reconstructed frames. Fig. 9 illustrates a notable distinction: While other event-to-video reconstruction methods often struggle with flow inference, leading to extensive areas of invalid flow estimation (manifested as white regions), our E2VIDiff achieves comprehensive flow coverage throughout the frame. This capability underscores the method's proficiency in capturing the motion.

4.5 Sample diversity

We present a qualitative analysis that highlights the diversity of the results produced by the proposed E2VIDiff compared to E2VID+ (Stoffregen et al, 2020) and ETNet (Weng et al, 2021) in Figs.1 and 10^4. Fig. 1 captures an outdoor driving scenario along a street. Unlike E2VID+ (Stoffregen et al, 2020) and ETNet (Weng et al, 2021), which tend to produce reconstructions with limited variability, the proposed E2VIDiff excels in generating four distinct outcomes from the same input, where each version reflects unique interpretations of environmental lighting and scene details, thereby illustrating E2VIDiff's adeptness at navigating the uncertainties inherent in real-world scenarios. Fig. 10 shows a car driving along a mountain road. Here, rapid movement and

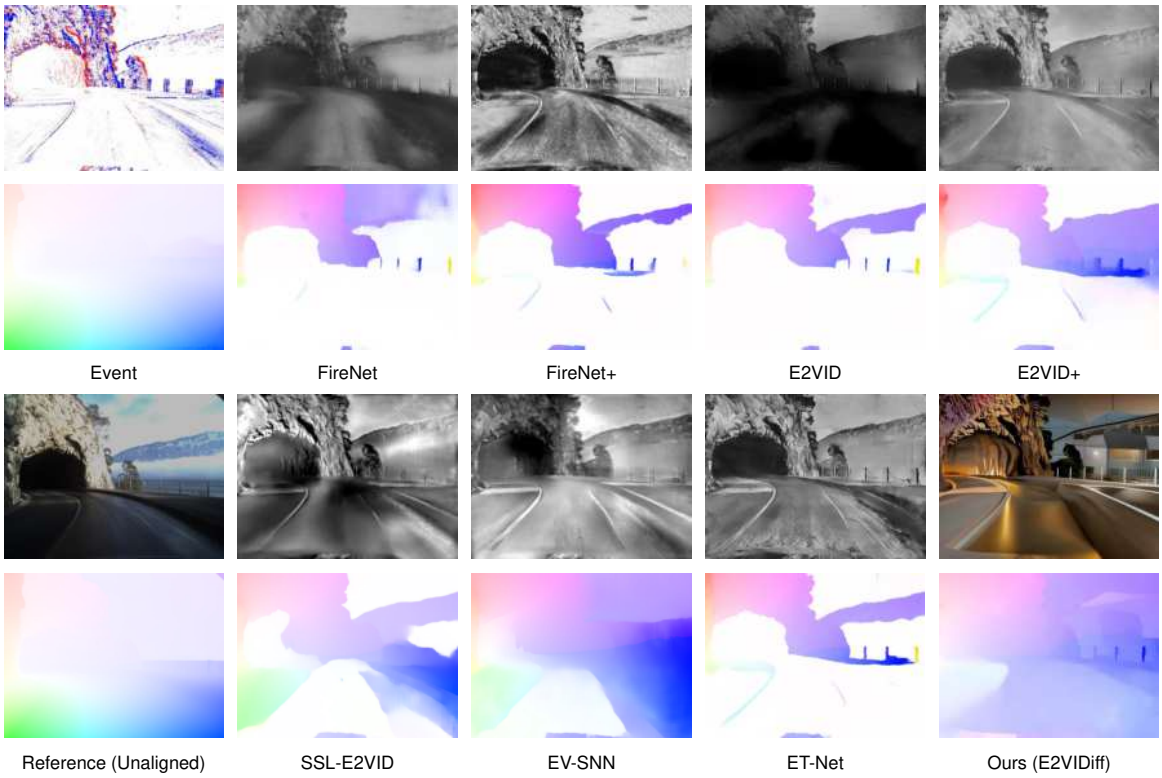


Fig. 9 Results of optical flow estimation using events-to-video reconstructions as an intermediary, allowing the direct application of conventional computer vision algorithms to event data. We apply E-RAFT (Gehrig et al, 2021b) to event data, with comparative results from RAFT (Teed and Deng, 2020) on unaligned reference and frames reconstructed by different methods. Please refer to the caption of Fig. 4 for a complete list of the compared methods.

challenging terrain test the temporal consistency and motion interpretation capabilities of the proposed E2VIDiff. While E2VID+ (Stoffregen et al, 2020) and ETNet (Weng et al, 2021) manage to capture the scene accurately, they offer a singular, constrained view. The proposed E2VIDiff, on the other hand, produces four reconstructions, each articulating different facets of the car's movement and the terrain's roughness. This demonstrates E2VIDiff's exceptional capability to interpret complex motion and terrain dynamics, offering diverse perspectives on fast-motion scenes.

5 Conclusion

In this work, we introduce E2VIDiff, leveraging pretrained diffusion models as image and video priors for perceptual events-to-video reconstruction. It tackles the inherent ambiguity in initial conditions and significantly improves the perceptual quality of the reconstructed frames. Through the development of an efficient encoder and integration module, we bridge the gap between pretrained diffusion models and events. Furthermore, the introduction of an event-guided sampling mechanism is instrumental in achieving higher fidelity in events-to-video reconstructions. These facilitate the generation of faithful, photorealistic, and chromatic reconstructions from achromatic events, marking an advancement in events-to-video reconstruction tasks.

Although the proposed E2VIDiff has shown remarkable performance in reconstructing video from events in scenarios characterized by high dynamic range and rapid motion (refer to our results shown in Figs. 5, 8, and 9), it falls short in generating high frame rate videos with subtle motion due to the scarcity of such data in the currently publicly accessible large scale datasets for training. Additionally, constraints within the motion priors in the pretrained models pose challenges in generating complex motions, such as intricate sports movements or the movement of uncommon objects. These hurdles can be overcome with advances in open source video foundation models that possess more robust motion priors, such as SORA (Brooks et al, 2024). Like other diffusion-based video generation methods, the proposed method also encounters the drawback of extended inference time, a challenge that could be addressed with more efficient sampling strategies. Furthermore, while the proposed method aims to harness the capabilities of existing video diffusion models for event-to-video reconstruction, it is necessary to recognize the potential misuse of such technology in generating deceptive content.

Fig. 10 Qualitative results demonstrating our method's diversity versus E2VID+ (Stoffregen et al, 2020) (the second column) and ETNet (Weng et al, 2021) (the third column). From the input achromatic event stream (first column), our approach yields diverse and plausible chromatic frame reconstructions (final four columns), each depicting distinct aspects of vehicular motion and terrain texture.

Acknowledgement

This work was supported by the National Science and Technology Major Project (Grant No. 2021ZD0109803), the National Natural Science Foundation of China (Grant No. 62302019, 62136001, and 62088102), and the China Postdoctoral Science Foundation (Grant No. 2022M720236).

References

Bain M, Nagrani A, Varol G, Zisserman A (2021) Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In: Proc. of International Conference on Computer Vision

Bardow P, Davison AJ, Leutenegger S (2016) Simultaneous Optical Flow and Intensity Estimation From an Event Camera. In: Proc. of Computer Vision and Pattern Recognition

Blattmann A, Rombach R, Ling H, Dockhorn T, Kim SW, Fidler S, Kreis K (2023) Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In: Proc. of Computer Vision and Pattern Recognition

Brooks T, Peebles B, Homes C, DePue W, Guo Y, Jing L, Schnurr D, Taylor J, Luhman T, Luhman E, Ng CWY, Wang R, Ramesh A (2024) Video Generation Models as World Simulators. Tech. rep., OpenAI, URL https://openai.com/research/video-generation-models-as-world-simulators

Cadena PRG, Qian Y, Wang C, Yang M (2021) SPADE-E2VID: Spatially-Adaptive Denormalization for Event-Based Video Reconstruction. IEEE Transactions on Image Processing 30:2488–2500

Carreira J, Zisserman A (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: Proc. of Computer Vision and Pattern Recognition

Choi J, Kim S, Jeong Y, Gwon Y, Yoon S (2021) ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In: Proc. of International Conference on Computer Vision

Dhariwal P, Nichol AQ (2021) Diffusion Models Beat GANs on Image Synthesis. In: Advances in Neural Information Processing Systems

Esser P, Rombach R, Ommer B (2021) Taming Transformers for High-Resolution Image Synthesis. In: Proc. of Computer Vision and Pattern Recognition

Gallego G, Delbrück T, Orchard G, Bartolozzi C, Taba B, Censi A, Leutenegger S, Davison AJ, Conradt J, Daniilidis K, Scaramuzza D (2022) Event-Based Vision: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(1):154–180

Gehrig D, Rebecq H, Gallego G, Scaramuzza D (2020) EKLT: Asynchronous Photometric Feature Tracking Using Events and Frames. International Journal of Computer Vision 128(3):601–618

Gehrig M, Aarents W, Gehrig D, Scaramuzza D (2021a) DSEC: A Stereo Event Camera Dataset for Driving Scenarios. IEEE Robotics and Automation Letters 6(3):4947–4954

Gehrig M, Millhäusler M, Gehrig D, Scaramuzza D (2021b) E-RAFT: Dense Optical Flow from Event Cameras. In: Proc. of International Conference on 3D Vision (3DV)

Guo Y, Yang C, Rao A, Liang Z, Wang Y, Qiao Y, Agrawala M, Lin D, Dai B (2023) AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In: International Conference on Learning Representations

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems

Kawar B, Vaksman G, Elad M (2021) SNIPS: Solving Noisy Inverse Problems Stochastically. In: Advances in Neural Information Processing Systems

Khachatryan L, Movsisyan A, Tadevosyan V, Henschel R, Wang Z, Navasardyan S, Shi H (2023) Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. In: Proc. of International Conference on Computer Vision

Kim H, Handa A, Benosman R, Ieng S, Davison AJ (2014) Simultaneous Mosaicing and Tracking with an Event Camera. In: Proc. of British Machine Vision Conference

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, et al (2023) Segment Anything. arXiv preprint arXiv:230402643

Lichtsteiner P, Posch C, Delbruck T (2008) A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits 43(2):566–576

Liu DC, Nocedal J (1989) On the Limited Memory BFGS Method for Large Scale Optimization. Mathematical Programming 45(1):503–528

Luo Z, Chen D, Zhang Y, Huang Y, Wang L, Shen Y, Zhao D, Zhou J, Tan T (2023) VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. In: Proc. of Computer Vision and Pattern Recognition

Mitrokhin A, Fermüller C, Parameshwara C, Aloimonos Y (2018) Event-Based Moving Object Detection and Tracking. In: IEEE/RSJ International Conference on Intelligent Robots and Systems

Mostafavi M, Wang L, Yoon KJ (2021) Learning to Reconstruct HDR Images from Events, with Applications to Depth and Flow Prediction. International Journal of Computer Vision 129(4):900–920

Mueggler E, Rebecq H, Gallego G, Delbrück T, Scaramuzza D (2017) The Event-Camera Dataset and Simulator: Event-Based Data for Pose Estimation, Visual Odometry, and SLAM. International Journal of Robotics Research 36(2):142–149

Munda G, Reinbacher C, Pock T (2018) Real-Time Intensity-Image Reconstruction for Event Cameras Using Manifold Regularisation. International Journal of Computer Vision 126(12):1381–1393

Nichol AQ, Dhariwal P, Ramesh A, Shyam P, Mishkin P, Mcgrew B, Sutskever I, Chen M (2022) GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In: Proc. of International Conference on Machine Learning

Paredes-Valles F, de Croon GCHE (2021) Back to Event Basics: Self-Supervised Learning of Image Reconstruction for Event Cameras via Photometric Constancy. In: Proc. of Computer Vision and Pattern Recognition

Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:220406125

Rebecq H, Gallego G, Mueggler E, Scaramuzza D (2018) EMVS: Event-Based Multi-View Stereo—3D Reconstruction with an Event Camera in Real-Time. International Journal of Computer Vision 126(12):1394–1414

Rebecq H, Ranftl R, Koltun V, Scaramuzza D (2019) Events-To-Video: Bringing Modern Computer Vision to Event Cameras. In: Proc. of Computer Vision and Pattern Recognition

Rebecq H, Ranftl R, Koltun V, Scaramuzza D (2021) High Speed and High Dynamic Range Video with an Event Camera. IEEE Transactions on Pattern Analysis and Machine Intelligence 43(6):1964–1980

Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-Resolution Image Synthesis With Latent Diffusion Models. In: Proc. of Computer Vision and Pattern Recognition

Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T, Ho J, Fleet DJ, Norouzi M (2022) Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In: Advances in Neural Information Processing Systems, vol 35

Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X, Chen X (2016) Improved Techniques for Training GANs. In: Advances in Neural Information Processing Systems

Scheerlinck C, Barnes N, Mahony R (2019a) Asynchronous Spatial Image Convolutions for Event Cameras. IEEE Robotics and Automation Letters 4(2):816–822

Scheerlinck C, Barnes N, Mahony R (2019b) Continuous-Time Intensity Estimation Using Event Cameras. In: Proc. of Asian Conference on Computer Vision

Scheerlinck C, Rebecq H, Gehrig D, Barnes N, Mahony RE, Scaramuzza D (2020) Fast Image Reconstruction with an Event Camera. In: Proc. of Winter Conference on Applications of Computer Vision

Song J, Meng C, Ermon S (2020) Denoising Diffusion Implicit Models. In: International Conference on Learning Representations

Song Y, Shen L, Xing L, Ermon S (2022) Solving Inverse Problems in Medical Imaging with Score-Based Generative Models. In: International Conference on Learning Representations

Stoffregen T, Scheerlinck C, Scaramuzza D, Drummond T, Barnes N, Kleeman L, Mahony R (2020) Reducing the Sim-to-Real Gap for Event Cameras. In: Proc. of European Conference on Computer Vision

Teed Z, Deng J (2020) RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In: Proc. of European Conference on Computer Vision

Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning Spatiotemporal Features With 3D Convolutional Networks. In: Proc. of International Conference on Computer Vision

Unterthiner T, van Steenkiste S, Kurach K, Marinier R, Michalski M, Gelly S (2019) FVD: A New Metric for Video Generation. In: ICLR 2019 Workshop on Deep Generative Models for Highly Structured Data

Vidal AR, Rebecq H, Horstschaefer T, Scaramuzza D (2018) Ultimate SLAM? Combining Events, Images, and IMU for Robust Visual SLAM in HDR and High-Speed Scenarios. IEEE Robotics and Automation Letters 3(2):994–1001

Voleti V, Jolicoeur-Martineau A, Pal C (2022) MCVD: Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation. In: Advances in Neural Information Processing Systems

Wang L, I SMM, Ho YS, Yoon KJ (2019) Event-Based High Dynamic Range Image and Very High Frame Rate Video Generation Using Conditional Generative Adversarial Networks. In: Proc. of Computer Vision and Pattern Recognition

Wang Y, Yu J, Zhang J (2022) Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. In: International Conference on Learning Representations

Weng W, Zhang Y, Xiong Z (2021) Event-Based Video Reconstruction Using Transformer. In: Proc. of International Conference on Computer Vision

Wu JZ, Ge Y, Wang X, Lei SW, Gu Y, Shi Y, Hsu W, Shan Y, Qie X, Shou MZ (2023) Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. In: Proc. of International Conference on Computer Vision

Xu X, Wang Z, Zhang E, Wang K, Shi H (2023) Versatile Diffusion: Text, Images and Variations All in One Diffusion Model. In: Proc. of Computer Vision and Pattern Recognition

Yu S, Sohn K, Kim S, Shin J (2023) Video Probabilistic Diffusion Models in Projected Latent Space. In: Proc. of Computer Vision and Pattern Recognition

Zhang H, Zhang L, Dai Y, Li H, Koniusz P (2022) Event-guided Multi-patch Network with Self-supervision for Non-uniform Motion Deblurring. International Journal of Computer Vision

Zhang L, Rao A, Agrawala M (2023) Adding Conditional Control to Text-to-Image Diffusion Models. In: Proc. of International Conference on Computer Vision

Zhou C, Teng M, Han J, Liang J, Xu C, Cao G, Shi B (2023) Deblurring Low-Light Images with Events. International Journal of Computer Vision 131(5):1284–1298

Zhu AZ, Thakur D, Ozaslan T, Pfrommer B, Kumar V, Daniilidis K (2018) The Multi Vehicle Stereo Event Camera Dataset: An Event Camera Dataset for 3D Perception. IEEE Robotics and Automation Letters 3(3):2032–2039

Zhu L, Wang X, Chang Y, Li J, Huang T, Tian Y (2022) Event-Based Video Reconstruction via Potential-Assisted Spiking Neural Network. In: Proc. of Computer Vision and Pattern Recognition

Zou Y, Zheng Y, Takatani T, Fu Y (2021) Learning To Reconstruct High Speed and High Dynamic Range Videos From Events. In: Proc. of Computer Vision and Pattern Recognition