# Diffusion-Based Extreme High-speed Scenes Reconstruction with the Complementary Vision Sensor

Yapeng Meng[1,†], Yihan Lin[1,2,†,‡], Taoyi Wang[1], Yuguo Chen[1], Lijian Wang[1], Rong Zhao[1,*]
[1]Department of Precision Instrument, Tsinghua University, Beijing, China,
[2]Pen-Tung Sah Institute of Micro-Nano Science and Technology, Xiamen University, Xiamen, China
{myp23, wangty23, cyg22, wlj24}@mails.tsinghua.edu.cn, linyh@xmu.edu.cn, r_zhao@tsinghua.edu.cn

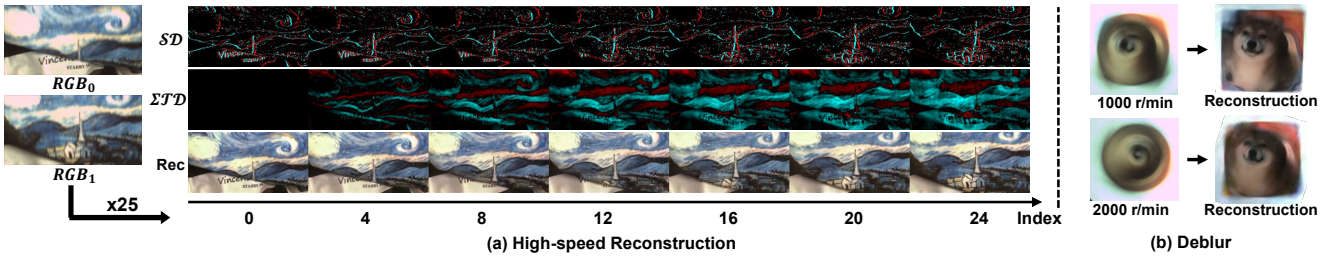


Figure 1. (a) A segment of real-captured data by the complementary vision sensor, Tianmouc [46], which includes adjacent 30 FPS RGB frames, spatial difference ($\mathcal{SD}$) and accumulated temporal difference ($\mathcal{TD}$) between them. We achieve accurate, colorful, high-speed, and blur-free video reconstruction. (b) Our method can also achieve reconstruction even in extremely blurry real-world scenarios.

## Abstract

*Recording and reconstructing high-speed scenes poses a significant challenge. While high-speed cameras can capture fine temporal details, their extremely high bandwidth demands make continuous recording unsustainable. Conversely, traditional RGB cameras, typically operating at 30 FPS, rely on frame interpolation to synthesize high-speed motion, often introducing artifacts and motion blur. Human visual system inspired sensors, like event cameras, offer high-speed sparse temporal or spatial variation data, partially alleviating these issues. However, existing methods still suffer from RGB blur, temporal aliasing, and loss of event information. To overcome these challenges, we leverage a novel complementary vision sensor, Tianmouc, which outputs high-speed, multi-bit, sparse spatio-temporal difference information with RGB frames. Building on this unique sensing modality, we introduce a* ***C****ascaded* ***Bi****-directional* ***R****ecurrent* ***D****iffusion* ***M****odel (CBRDM) that achieves accurate, sharp, color-rich video frames reconstruction. Our method outperforms state-of-the-art RGB interpolation algorithms in quantitative evaluations and surpasses event-based methods in real-world comparisons. Code and dataset are at* https://github.com/Tianmouc/GenRec.

## 1. Introduction

Recording and reconstructing extreme high-speed motion scenarios is crucial for autonomous driving [44], sports analytics [30], industrial inspection [38], and scientific research [6]. While high-speed cameras achieve exceptionally high frame rates, they suffer from high bandwidth demands and substantial storage costs, making prolonged continuous recording impractical [35]. Therefore, there is a growing interest in using low-frame-rate cameras to capture data and accurately reconstruct sharp, high-speed scenes with post-processing algorithms. However, the low-frame-rate cameras struggle to capture real-world non-linear motion and illumination variations, while motion blur further degrades the quality of the data. Given this ill-conditioned nature, existing RGB-based approaches such as video deblur [33] and video frame interpolation [17, 25, 47] often produce blurring, distortions, and artifacts in high-speed scenarios.

The advent of neuromorphic sensors (e.g., DVS [23], DAVIS [2], and spiking cameras [15]) provides an innovative solution. While maintaining similar bandwidth and storage costs to those of conventional RGB cameras, neuromorphic sensors feature high-speed data pathways to capture scene dynamics effectively. For example, event cameras asynchronously measure brightness changes at the pixel level, and algorithms facilitating event and the RGB information [40] demonstrate superior interpolation performance in high-speed scenarios compared to RGB-only interpolation methods. However, event cameras struggle to

[†]These authors contributed equally to this work
[‡]This work was performed while the author was at Tsinghua University
[*]Corresponding author

capture object edges parallel to the motion [11] and suffer from event rate saturation [7] and timestamp distortion [16] under rapid changes, leading to information loss. In addition, achieving spatial alignment and temporal synchronization between event cameras and RGB cameras is challenging, often requiring meticulous beam-splitting optical design [18, 26, 40] and hardware synchronization. Some multi-pathway vision sensors integrating event and RGB have addressed [2, 9, 21] this systemic issue with low bandwidth cost, providing more promising hardware support for this field. However, they are still constrained by the inherent limitations of the event modality as mentioned above.

In this work, we employ a novel dual-pathway complementary vision sensor (CVS), Tianmouc [46], which can output a 30 FPS RGB video stream (cognition-oriented pathway, COP) while simultaneously providing a high-speed ($\geq$757 FPS) sparse and multi-bit spatio-temporal differential data stream (action-oriented pathway, AOP). Both pathways are hardware-synchronized in time and space, providing a complementary mechanism that offers the essential information necessary for high-speed reconstruction. The spatial difference ($\mathcal{SD}$) data captures rich edge contours and texture details, which significantly enhance the deblurring of RGB frames and enable precise object localization. Meanwhile, the temporal difference ($\mathcal{TD}$) data accurately records illumination changes and the high-speed moving objects within the scene. Using $\mathcal{SD}$ and $\mathcal{TD}$, we achieve a fast, complete, stable, and high-precision description of the scene while maintaining sparsity and low bandwidth, establishing a solid foundation for low-bandwidth recording and high-fidelity reconstruction in challenging high-speed scenarios. However, due to differences in data distribution and spectral characteristics, reconstructing the three data modalities into a visible high-speed RGB video stream remains a challenge.

Based on the CVS, we have proposed a Cascade Bi-directional Recurrent Diffusion Model (CBRDM) that facilitates high-fidelity reconstruction of sharp, high-speed, and color-rich video streams under high-speed conditions. To effectively leverage the conditional information provided by the two complementary data pathways, we propose a diffusion-based two-stage generative reconstruction framework. Non-generative approaches often rely on linear or overly simplistic motion assumptions, limiting their effectiveness compared to generative models. In cases of severe motion blur, complex motion patterns, or rapid light changes between reference frames, these methods struggle to produce plausible reconstructions [34, 39]. In contrast, a key advantage of generative models is their ability to produce diverse samples rather than regressing toward average values, making them well-suited for reconstructing extreme high-speed scenes. Reconstructing dense RGB frames requires generative models to enforce temporal consistency and retain memory. We enhance standard diffusion models with a bi-directional recurrent mechanism in the encoder, allowing intermediate frames to integrate forward and backward feature maps during encoding. This improves temporal awareness, effectively mitigating color loss and confusion. Traditional diffusion models face challenges such as high memory consumption and slow convergence at high resolutions, while latent diffusion models, despite reducing memory usage, struggle to recover fine details accurately. To address this, we adopt a two-stage cascaded architecture in CBRDM, where a coarse reconstruction is first performed at low resolution, followed by a super-resolution stage under conditional constraints to restore the original resolution. This approach reduces computational costs and simplifies training while maintaining high reconstruction fidelity.

Collecting large-scale data for a novel sensor also poses a significant challenge. We employ a vision chip characterization method [27] to convert high-speed RGB video datasets into the CVS data format. This method utilizes a Digital micro-mirror device (DMD) chip and corresponding optical paths to project light onto the sensor's pixels. Additionally, we test our algorithm's performance on real-world captured scenes. Both benchmark evaluation metrics and visual results from real-captured data demonstrate that our method achieves state-of-the-art reconstruction performance.

The contribution of our paper is summarized as follows:

- We propose a novel solution that combines the dual-pathway CVS with a diffusion-based video reconstruction algorithm, CBRDM, to enable the recording and reconstruction of extreme high-speed motion scenarios. This method effectively utilizes 30 FPS RGB data and high-speed spatio-temporal difference data to achieve consistent, blur-free, high-speed ($\geq$757 FPS) video reconstruction, thereby avoiding significant bandwidth and storage consumption.
- We propose a Bi-directional Recurrent UNet Block for diffusion models, enabling the network to achieve temporal awareness. This design ensures temporal consistency in video reconstruction, effectively preventing color loss and confusion in intermediate frames.
- Our method achieves significantly superior performance across multiple high-speed RGB datasets. Without requiring any fine-tuning, our model can be directly applied to real-world captured data, demonstrating state-of-the-art practical reconstruction results.

## 2. Related Work

**Joint Video Deblurring and Interpolation.** When reconstructing extreme high-speed scenes, on the one hand, it is necessary to interpolate intermediate frames between low-temporal-resolution RGB frames; on the other hand, due to limitations in sensor shutter speed and exposure time, the

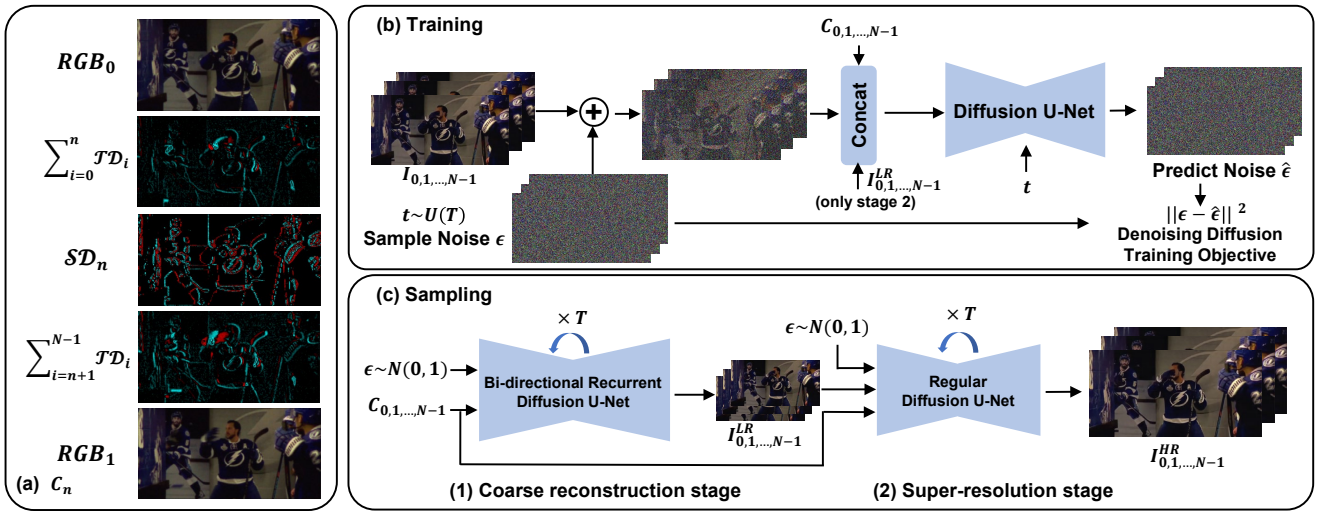


Figure 2. An overview of our proposed method. (a) Visualization of the condition input. (b) A single training step of our denoising diffusion model, where we concatenate the noisy sample and the condition input as the UNet input. (c) The sampling process begins with coarse reconstruction at a low resolution, followed by super-resolution to restore the original resolution.

captured RGB frames are often blurred, necessitating the joint consideration of deblurring and denoising tasks.

Some early methods [24] employed cascaded frameworks to achieve deblurring and frame interpolation across different stages. Zhang et al. [50] proposed a self-supervised unified framework for event cameras, exploiting the mutual constraints among blurry frames, sharp latent frames, and event streams to address both tasks. Sun et al. [39] introduced a bidirectional recurrent network to implement event-based frame interpolation that performs deblurring ad hoc. However, most joint deblurring and interpolation models based on event cameras explicitly rely on the exposure time of RGB cameras, utilizing the event streams captured during the exposure period for further processing. Yang et al. [45], however, identified the temporal discrepancy between the actual occurrence of changes and the corresponding timestamps assigned by the sensor. This delay introduces errors to algorithmic frameworks dependent on exposure time. In contrast, the CVS leverages its high-speed $\mathcal{SD}$ pathway to directly obtain clear edge contour information without reliance on exposure time.

**Diffusion Model in low-level vision.** Diffusion Models [12, 37] model the image generation process as progressive denoising a Gaussian noise image. Compared to other generative models such as GANs [8] and VAEs [19], diffusion models have achieved state-of-the-art results in tasks such as text-to-image generation [32], text-to-video generation [14], and conditional image/video generation [10, 48]. Recently, they have also been widely applied to low-level computer vision tasks, including super-resolution [42], deblurring [4, 31], and video frame interpolation [5, 17]. For ill-conditioned tasks with multiple solutions, using generative models instead of directly optimizing pixel-level losses is beneficial for achieving higher perceptual quality. Additionally, while adopting the approach of performing low-level vision tasks in latent space, as seen in text-image models, can help reduce the computational cost, it may struggle to handle fine details like small texts, faces, and patterns [42]. Therefore, our model performs diffusion directly in the pixel space while reducing computational cost through a two-stage cascading approach.

## 3. Methodology

### 3.1. Complementary Vision Sensor

The complementary vision sensor, Tianmouc [46], implements two pathways: the RGB pathway (30 FPS, 10-bit, $\mathbf{RGB} \in \mathbb{R}^{H\times W\times 3}$), and the high-speed spatio-temporal difference pathway, which supports an adaptive frame rate (757–10,000 FPS) and pixel precision ranging from $\pm$1-bit to $\pm$7-bit. The definition of the temporal difference ($\mathcal{TD} \in \mathbb{R}^{H\times W\times 1}$) and spatial difference ($\mathcal{SD} \in \mathbb{R}^{H\times W\times 2}$) pathways can be expressed as follows:

$$\begin{aligned} \mathcal{TD} &= \nabla_t \mathbf{I}, \\ \mathcal{SD} &= Concat(\nabla_{+45^\circ}\mathbf{I}; \nabla_{-45^\circ}\mathbf{I}), \end{aligned} \tag{1}$$

where $\mathbf{I} \in \mathbb{R}^{H\times W}$ represents the light intensity, $\nabla_t$ denotes the temporal gradient, $\nabla_{\pm 45^\circ}$ represent spatial gradients computed along ±45° directions (due to the cross-pixel design of the CVS circuit), and $Concat(;)$ denotes concatenating the frames along the channel dimension.

### 3.2. Problem Formulations

Given two blurry, low-temporal-resolution RGB keyframes, $\mathbf{RGB}_0$ and $\mathbf{RGB}_1$, along with $N$ high-temporal-resolution spatial difference frames $\mathcal{SD}_n$ and temporal difference frames $\mathcal{TD}_n$ for $n \in \{0, 1, \ldots, N-1\}$, our goal is to reconstruct a sequence of clear, accurate, and color-rich high-temporal-resolution RGB frames, $\mathbf{F}_n$. Note that $(\mathcal{TD}_0, \mathcal{SD}_0)$ and $(\mathcal{TD}_{N-1}, \mathcal{SD}_{N-1})$ are synchronized with $\mathbf{RGB}_0$ and $\mathbf{RGB}_1$, respectively.

### 3.3. Algorithm Overview

As shown in Fig. 2, we formulate video frame reconstruction as a conditional denoising diffusion problem and design a two-stage denoising network to model the conditional distribution $D(\mathbf{F}|\mathbf{C})$ over $N$ video frames, $\mathbf{F} \in \mathbb{R}^{N\times H\times W\times 3}$. The overall conditioning input, $\mathbf{C} \in \mathbb{R}^{N\times H\times W\times 10}$, is constructed by concatenating the frame-wise conditions $\mathbf{C}_n$ along the temporal dimension:

$$\mathbf{C} = \text{Concat}(\mathbf{C}_0; \mathbf{C}_1; \ldots; \mathbf{C}_{N-1}), \tag{2}$$

where each frame-wise condition $\mathbf{C}_n \in \mathbb{R}^{H\times W\times 10}$ is defined as:

$$\mathbf{C}_n = \text{Concat}(\mathbf{RGB}_0; \sum_{i=0}^{n} \mathcal{TD}_i; \mathcal{SD}_n; \sum_{i=n+1}^{N-1} \mathcal{TD}_i; \mathbf{RGB}_1). \tag{3}$$

We propose a two-stage cascaded denoising network (detailed in Sec. 3.4). In the coarse reconstruction stage, we introduce a Bi-directional Recurrent Denoising UNet (detailed in Sec. 3.5) to simultaneously reconstruct $N$ frames under the low-resolution condition $\mathbf{C}^{LR}$ (bilinear downsampled by $\mathbf{C}$), generating low-resolution $\mathbf{F}^{LR}$. In the super-resolution stage, the network refines the results at the original resolution using a spatial convolutional self-attention UNet, conditioned on $\mathbf{C}$ and the first-stage outputs $\mathbf{F}^{LR}$, to generate the final result $\mathbf{F}$.

### 3.4. Cascaded Diffusion Model

**Coarse reconstruction stage.** To enable the network to have a global receptive field and reduce computational complexity and memory usage, we draw inspiration from the cascaded diffusion model strategy used in video generation models [13]. In this stage, $N$ consecutive frames $\mathbf{F}^{LR}$ are reconstructed simultaneously at a low resolution ($48 \times 96$), conditioned on the downsampled input $\mathbf{C}^{LR}$. The forward process involves progressively adding Gaussian noise to $\mathbf{F}^{LR}$ over $T$ steps, as defined below:

$$\mathbf{F}_t^{LR} = \sqrt{\bar{\alpha}_t}\mathbf{F}_0^{LR} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{4}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$, $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, and $\{\beta_1, \ldots, \beta_T\}$ represents the variance schedule. The reverse process denoises $\mathbf{F}_t^{LR}$ using the learned model $\mathcal{M}_{\theta_1}(\mathbf{F}_t^{LR}, \mathbf{C}^{LR}, t)$ to reconstruct $\mathbf{F}_0^{LR}$.

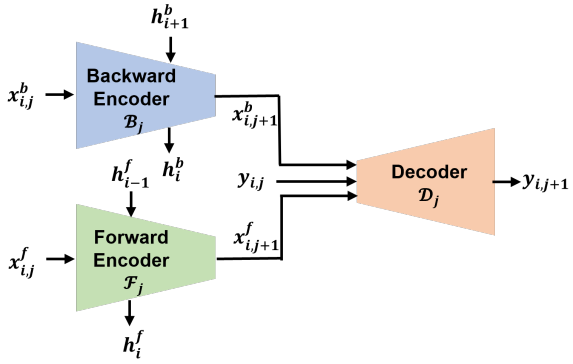


Figure 3. Illustration of the proposed Bi-directional Recurrent Block. The figure depicts the processing of the $i^{th}$ sample by the $j^{th}$ convolutional block.

At training time, parameters $\theta_1$ are updated by minimizing the denoising diffusion objective functions [12] as follows:

$$\mathcal{L} = \mathbb{E}_{\mathbf{F}_0^{LR}, \boldsymbol{\epsilon}\sim\mathcal{N}(0,I), t\sim\mathcal{U}(T)} \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}\|_2^2. \tag{5}$$

**Super-resolution stage.** Each frame is refined individually at the sensor's equivalent resolution ($320 \times 640$) using the first stage output $\mathbf{F}_n^{LR}$ and the frame-specific condition $\mathbf{C}_n$. The forward process is similar to Eq. 4:

$$\mathbf{F}_{n,t} = \sqrt{\bar{\alpha}_t}\mathbf{F}_{n,0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}. \tag{6}$$

The training objective follows Eq. (5), with parameters $\theta_2$ optimized using the estimated noise $\hat{\epsilon} = \mathcal{M}_{\theta_2}(\text{UP}(\mathbf{F}_n^{LR}), \mathbf{F}_{n,t}, \mathbf{C}_n, t)$, UP denotes upsampling by bilinear interpolation.

**Inference stage.** First, $\mathbf{F}^{LR}$ is reconstructed from a normally distributed variable $\mathbf{F}_T^{LR}$ by iteratively applying the learned denoiser $\mathcal{M}_{\theta_1}(\mathbf{F}_t^{LR}, \mathbf{C}^{LR}, t)$. Subsequently, using a similar process, each frame $\mathbf{F}_n$ of $\mathbf{F}$ is obtained iteratively from normally distributed Gaussian noise by the super-resolution stage denoiser $\mathcal{M}_{\theta_2}(\text{UP}(\mathbf{F}_n^{LR}), \mathbf{F}_{n,t}, \mathbf{C}_n, t)$.

### 3.5. Bi-directional Recurrent Denoising UNet Block

Our task aims at reconstructing multiple video frames between two RGB frames. Given the significant differences between the two RGB frames in high-speed scenarios, ensuring the continuity and consistency of the intermediate frames requires the model to have temporal awareness. However, incorporating temporal blocks, such as those in the video diffusion model [1], substantially escalates both the parameter count and computational burden of the network. Additionally, temporal convolution layers suffer from a limited temporal receptive field, requiring multiple convolutional layers for critical information from edge frames to propagate to the center frames. Meanwhile, the temporal attention mechanism imposes constraints on the number

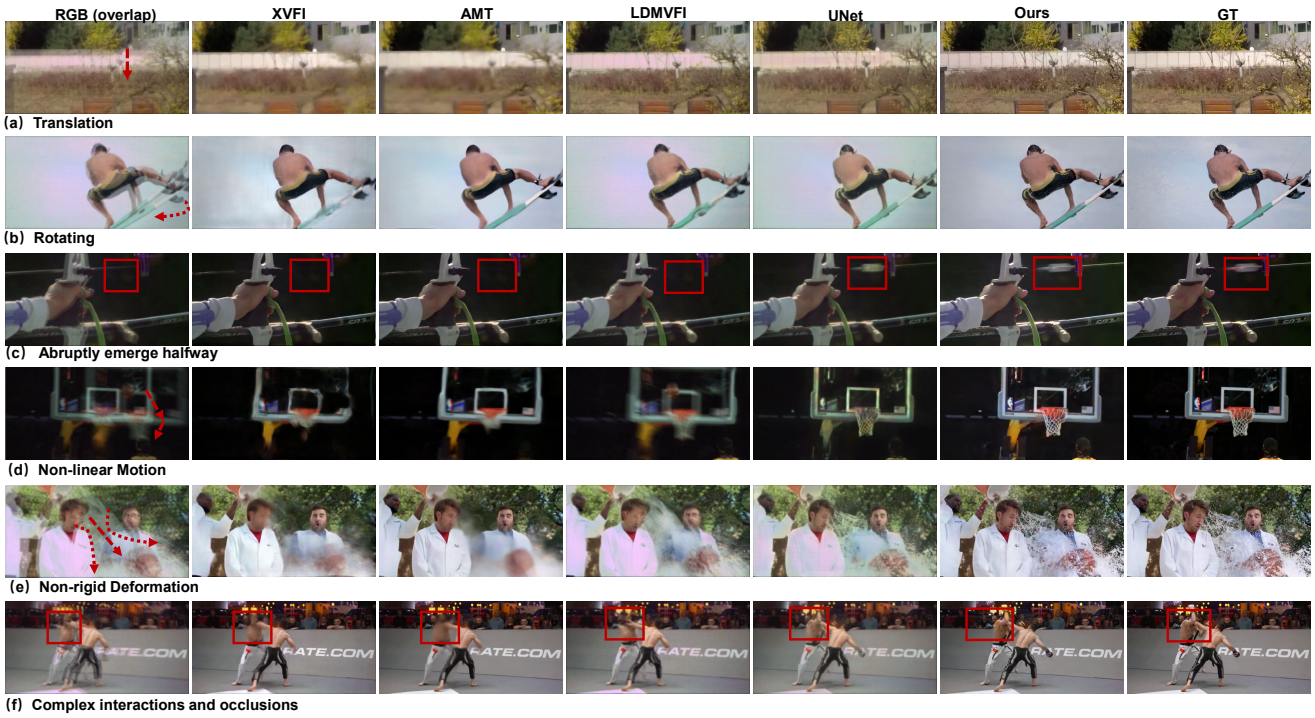


Figure 4. Comparison results with state-of-the-art methods. We show six typical categories: translation, rotation, abruptly emerging, non-linear motion, non-rigid deformation, and complex interactions with occlusion.

of frames generated during training, making it challenging to adapt to varying frame-rate data. To address these limitations, we propose a Bi-directional Recurrent Denoising UNet Block (BRB, as shown in Fig. 3), where each UNet block in the downsampling stage acquires temporal information from both forward and backward feature maps, formulated as:

$$\hat{x}^b_{i,j+1}, h^b_i = \mathcal{B}_j(x^b_{i,j}, h^b_{i+1}), \tag{7}$$

$$\hat{x}^f_{i,j+1}, h^f_i = \mathcal{F}_j(x^f_{i,j}, h^f_{i-1}), \tag{8}$$

$$x^b_{i,j+1} = \text{Down}(\hat{x}^b_{i,j+1}), \tag{9}$$

$$x^f_{i,j+1} = \text{Down}(\hat{x}^f_{i,j+1}), \tag{10}$$

where $x$ denotes the feature map in the downsampling stage, $i$ denotes the temporal index, $j$ denotes the convolutional block index, $\mathcal{B}$ refers to the backward encoder (input start at the last time step $i = N - 1$), $\mathcal{F}$ refers to the forward encoder (input start at the first time step $i = 0$) and Down denotes the downsampled convolution layer. Specifically, we define $h^b_N = \mathbf{0}$ and $h^f_{-1} = \mathbf{0}$.

During the upsampling stage, each decoder block $\mathcal{D}_j$ progressively upsamples the features by concatenating the forward feature map $x^f_{i,j}$ and the backward feature map $x^b_{i,j}$ from the downsampling stage, along with the feature map $y_{i,j}$ from the last decoder block $\mathcal{D}_{j-1}$, formulated as:

$$\hat{y}_{i,j+1} = \mathcal{D}_j(\text{Concat}(x^b_{i,j}, x^f_{i,j}, y_{i,j})), \tag{11}$$

$$y_{i,j+1} = \text{Up}(\hat{y}_{i,j+1}), \tag{12}$$

where $y$ denotes the feature map in the upsampling stage and Up denotes the upsampling operation by nearest interpolation and convolution.

### 3.6. Implementation Details

**Network architecture.** In the coarse reconstruction stage, our Bi-directional Recurrent Denoising UNet processes $48 \times 96$ resolution inputs using four residual blocks with output channels of $[160, 320, 320, 640]$. Each block contains several convolutional layers and spatial self-attention layers. In the super-resolution stage, the Denoising UNet operates at $320 \times 640$ resolution with a similar but lighter structure. The self-attention is only applied at the smallest scale to reduce computational cost. Details are provided in our publicly available code.

**Training details.** We implement our approach using PyTorch [29] framework and the Diffusers [41] library. All parameters are trained from scratch. We adopt the DDIM [37] sampling strategy, training with 1000 denoising steps. During inference, the first-stage network uses 200 steps, while the second-stage network operates with 50 steps. The Adam [20] optimizer is employed with a learning rate

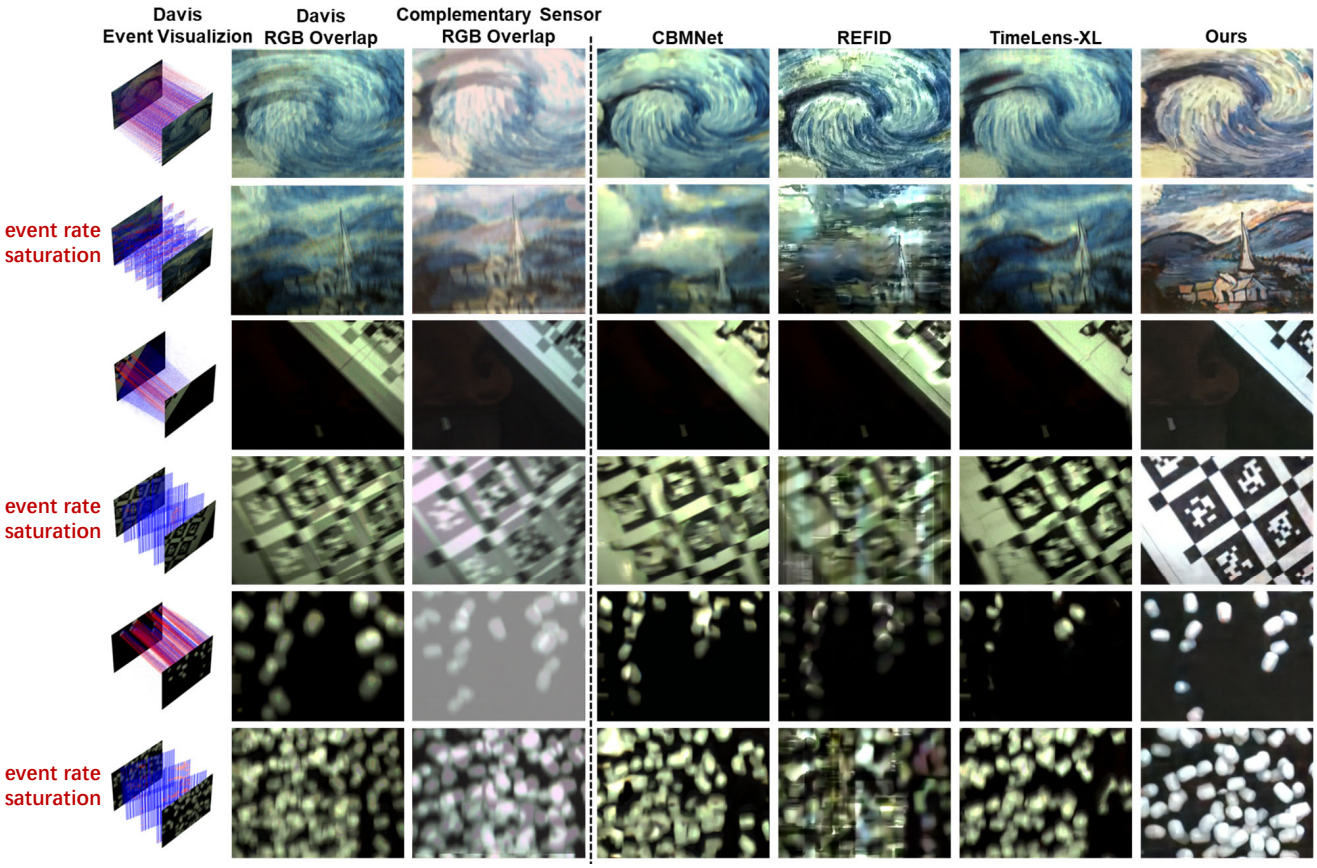


Figure 5. Results on real-captured data compared with event-based methods. When the scene changes rapidly, event rate saturation may occur, leading the algorithm to produce more severe errors.

decaying from 1e-5 to 1e-7 following an IterExponential schedule. The coarse reconstruction network is trained for 200K iterations, and the super-resolution network is trained for 100K iterations. We use 8 NVIDIA RTX 4090 GPUs for training and testing.

## 4. Experimental Analysis

### 4.1. Datasets

For a novel neuromorphic sensor, one of the major challenges in algorithm training and evaluation is acquiring a large volume of real-world data. Unlike the commonly used approach of combining simulators with real-world data captured by beam splitters for training, we employ a vision chip characterization method [27] that incorporates authentic sensor response, which utilizes a DMD chip and corresponding optical paths to control and project light onto the sensor's pixels.

Using this method, we convert high-speed RGB video datasets, including GoPro [28], X4K1000FPS [36], and SportsSlomo [3], into the CVS data format. Since the DMD-based data conversion method requires a continuous sequence of a specific number of images to generate complete RGB frames and $\mathcal{TD}, \mathcal{SD}$ frames between RGB (e.g., in multiples of 25 when the sensor operates at 757 FPS), we have re-partitioned these datasets into train, valid, and test sets (Parts of SportsSlomo for validing and X4K1000FPS only for training). Given the substantial data requirements of diffusion model training, we combine the training data from several datasets while maintaining separate testing on individual test sets. All comparison methods are trained using the same data strategy. Detailed descriptions of dataset processing and partitioning are provided in our supplementary materials.

### 4.2. Experiment Results

We compare our method with state-of-the-art video interpolation approaches, including the optical flow-based methods XVFI [36] and AMT [22], as well as the diffusion-based method LDMVFI [5]. Since event-based video interpolation methods utilize a different data format from the CVS, we further conduct a real-world comparison with these methods in Sec. 4.3. As we restructured the dataset and adopted a frame sampling strategy, extracting one frame

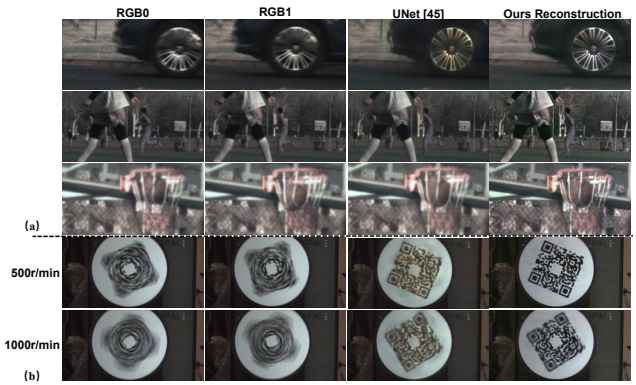


Figure 6. (a) Visualization results on real-captured data. (b) Testing the algorithm's performance on a high-speed rotating disk.

every 25 frames for interpolation, we retrained the comparison methods using their official implementations. The quantitative results are presented in Tab. 1, and the categorical visualization results are shown in Fig. 4. Given that our defined task involves handling blurry RGB frames, intermediate frame interpolation, and corrections to the camera ISP algorithm, which is different from the conventional video interpolation task, we further conduct an "Only Interpolation" experiment. In this setup, we use clear RGB frames as inputs and evaluate the interpolation performance of these methods independently. For evaluation, we employ three key metrics: pixel-level accuracy metric PSNR, structural similarity metric SSIM [43], and perpetual metric LPIPS [49]. In terms of evaluation metrics, our method significantly outperforms the compared approaches. When dealing with large-scale, non-linear motion and non-rigid deformation, as well as occlusions between objects, RGB-based methods struggle to achieve accurate scene reconstruction due to the lack of inter-frame spatio-temporal information. We also compare our method with the publicly available CVS reconstruction approach [46], an optical flow-based network built on a UNet backbone (abbreviated as "UNet" in figures throughout the paper), leveraging its pre-trained weights. The "UNet" method, which utilizes spatio-temporal difference data, performs well in reconstructing image structures but suffers from color loss and some blurring artifacts. In contrast, our method achieves the best visual quality.

### 4.3. Real-world Comparisons

We further conduct a real-world comparison between the CVS combined with our reconstruction algorithm and the event camera with the corresponding algorithms. Specifically, we utilize the DAVIS 346 camera for scene recording and compare our method with state-of-the-art event-based video interpolation methods CBMNet [18], TimeLens-XL [26], and the joint deblurring and interpolation method

Table 1. Quantitative Comparison with State-of-the-Art Methods.

| | GoPro [28] | | | SportsSlomo [3] | | |
|---|---|---|---|---|---|---|
| **Method** | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| XVFI [36] | 20.18 | 0.6028 | 0.3683 | 20.26 | 0.6691 | 0.3626 |
| XVFI† | 21.15 | 0.6127 | 0.3230 | 21.31 | 0.7014 | 0.2810 |
| AMT [22] | 21.74 | 0.6483 | 0.3248 | 22.56 | 0.7519 | 0.2673 |
| AMT† | 21.17 | 0.6157 | 0.3450 | 22.04 | 0.7305 | 0.2530 |
| LDMVFI [5] | 20.64 | 0.5723 | 0.3123 | 19.72 | 0.5930 | 0.3395 |
| LDMVFI† | 20.26 | 0.5675 | 0.2744 | 20.06 | 0.6643 | 0.2640 |
| Ours | **27.65** | **0.8986** | **0.1769** | **25.43** | **0.8252** | **0.1647** |

† Only Interpolation mode, detailed in Sec. 4.2.

REFID [39]. We employ the official network weights of these methods during inference. To ensure a consistent experimental setup, we carefully align the viewpoints of the CVS and the event camera and employ software-based synchronization to achieve approximate temporal alignment. Visualization results of the middle frame reconstruction are shown in Fig. 5. For each experimental scenario, we conduct tests at various motion speeds. When the scene changes rapidly, the DAVIS camera may experience event saturation, leading to reconstruction failure. In cases of minor scene changes, the DAVIS camera outputs remain stable; however, the comparison methods still suffer from blurring and artifacts, while our approach achieves superior visual quality.

### 4.4. Real World and Extreme Performance Testing

Our algorithm, despite being trained on DMD-based RGB-converted data, generalizes well to real-world captured data without any fine-tuning. The corresponding experimental results are shown in Fig. 6 (a). Furthermore, to evaluate the algorithm's extreme performance, we conducted experiments using a high-speed rotating QR code captured on a precisely controlled high-speed turntable. Fig. 6 (b) presents the results. At 500 r/min, the RGB image exhibits severe motion blur, making the QR code unreadable, whereas our algorithm still produces a clear reconstruction. When the speed increases to 1000 r/min, slight blurring appears at the edges of our reconstructed results, but the QR code remains recognizable on most devices—try scanning it! Additionally, we compared our method with the sensor's public reconstruction approach "UNet" [46], which produces noisier reconstructions and is more prone to color shifts.

### 4.5. Ablation Study

To comprehensively validate the necessity of dual-pathway CVS combined with our proposed diffusion-based generative reconstruction algorithm framework, we conduct ablation studies from two perspectives: input information and network structure. More ablation study results can be found in the supplementary material.


Figure 7. Visualization of the temporal awareness ablation experiment. TCAB refers to the temporal convolution and attention block used by [1]. BRB refers to our proposed bi-directional recurrent block. The highlighted red boxes indicate potential color loss and confusion in regions with large motions or complex textures.

Table 2. $\mathcal{TD}$/$\mathcal{SD}$ inputs ablation study on SportsSlomo dataset.

| $\mathcal{TD}$ | $\mathcal{SD}$ | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|---|
| × | × | 18.18 | 0.5651 | 0.2190 |
| × | ✓ | 21.98 | 0.7895 | 0.1238 |
| ✓ | × | 23.65 | 0.8403 | 0.1083 |
| ✓ | ✓ | **23.92** | **0.8512** | **0.0994** |

Table 3. Time awareness ablation study on SportsSlomo dataset.

| Temporal | Params | PSNR↑ | SSIM↑ | LPIPS↓ | PTC[1]↓ |
|---|---|---|---|---|---|
| × | 113.63M | 23.92 | 0.8512 | 0.0994 | 0.3674 |
| TCAB[2] | 249.88M | 24.12 | 0.8603 | 0.0897 | 0.3488 |
| BRB[3] | 166.18M | **24.36** | **0.8662** | **0.0806** | **0.3325** |

[1] PTC: Perceptual temporal consistency.
[2] TCAB: the Temporal Convolution and Attention Block [1].
[3] BRB: the Bi-directional Recurrent Block.

#### 4.5.1. Effectiveness of TD and SD as Condition Input

To validate the effectiveness of $\mathcal{TD}$ and $\mathcal{SD}$ information provided by the CVS, we performed an ablation study using a simple experimental baseline: a single-stage denoising model operating in a $48 \times 96$ pixel space, without any temporal awareness structure. We remove $\mathcal{TD}$ and $\mathcal{SD}$ input individually or jointly and evaluate the impact on various performance metrics. The results are reported in Tab. 2. Compared to removing only $\mathcal{TD}$ or $\mathcal{SD}$ inputs, full input results in an additional $19.7\%$ and $8.2\%$ reduction in LPIPS, respectively, demonstrating the contribution of spatio-temporal difference information provided by the CVS in high-quality reconstruction.

#### 4.5.2. Effectiveness of our network architecture

**Usage of the Bi-directional Recurrent Block.** To evaluate the effectiveness of our proposed BRB in enhancing the temporal awareness of the diffusion framework, we compare BRB with two alternatives: its removal and the insertion of temporal blocks (including temporal convolution and attention blocks, TCAB) as proposed in [1]. All methods use the same UNet architecture, number of layers, and channel configurations, and are trained to converge. We evaluate the perceptual temporal consistency by $\left\|\phi_i - \frac{1}{2}(\phi_{i-1} + \phi_{i+1})\right\|_1$, $\phi_i$ denotes VGG-16 feature maps at frame $i$. The evaluation metrics are presented in Tab. 3, and visualization results are shown in Fig. 7. The PSNR metric increases by 0.2 dB and 0.44 dB, and the perceptual metric LPIPS decreases by $9.8\%$ and $18.9\%$ when using TCAB and our proposed BRB, respectively. The visual results further demonstrate the importance of temporal awareness in preserving colors during large motion. Moreover, our BRB outperforms TCAB across all evaluation metrics while also reducing the network's parameter count.

## 5. Conclusion

To record and reconstruct extreme high-speed scenes, we employ a novel dual-pathway complementary vision sensor, Tianmouc, which captures 30 FPS RGB frames alongside high-speed (≥757 FPS) spatio-temporal difference frames. To achieve accurate, high-speed, sharp, and colorful video reconstruction, we design the CBRDM with a two-stage bi-directional recurrent diffusion block and cascade architecture. The CBRDM surpasses existing SOTA high-speed reconstruction algorithms on test datasets derived from GoPro and SportsSlomo. The high-quality data with real camera responses enables our model can be directly applied to real-world captured scenes. We also conducted real-world comparisons with event camera methods, showing that our sensor-data-algorithm framework delivers superior visual quality for high-speed imaging and reconstruction. Ablation studies further validate the effectiveness of the spatio-temporal difference data provided by the complementary vision sensor, as well as the proposed two-stage generative and temporal-aware network architecture.

**Acknowledgements.** This work is supported by the Brain Science and Brain-like Intelligence Technology-National Science and Technology Major Project 2021ZD0200300 and the Tsinghua University Initiative Scientific Research Program 20257020014.

## References


[1] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 4, 8

[2] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 $\mu$s latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 1, 2

[3] Jiaben Chen and Huaizu Jiang. Sportsslomo: A new benchmark and baselines for human-centric video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6475–6486, 2024. 6, 7

[4] Zheng Chen, Yulun Zhang, Ding Liu, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Hierarchical integration diffusion model for realistic image deblurring. *Advances in neural information processing systems*, 36, 2024. 3

[5] Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1472–1480, 2024. 3, 6, 7

[6] Daniel J Duke, Thomas Knast, Bhavraj Thethy, Luis Gisler, and Daniel Edgington-Mitchell. A low-cost high-speed cmos camera for scientific imaging. *Measurement Science and Technology*, 30(7):075403, 2019. 1

[7] Daniel Gehrig and Davide Scaramuzza. Are high-resolution event cameras really needed? *arXiv preprint arXiv:2203.14672*, 2022. 2

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[9] Menghan Guo, Shoushun Chen, Zhe Gao, Wenlei Yang, Peter Bartkovjak, Qing Qin, Xiaoqin Hu, Dahei Zhou, Masayuki Uchiyama, Yoshiharu Kudo, et al. A 3-wafer-stacked hybrid 15mpixel cis+ 1 mpixel evs with 4.6 gevent/s readout, in-pixel tdc and on-chip isp and esp function. In *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 90–92. IEEE, 2023. 2

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pages 330–348. Springer, 2025. 3

[11] Botao He, Ze Wang, Yuan Zhou, Jingxi Chen, Chahat Deep Singh, Haojia Li, Yuman Gao, Shaojie Shen, Kaiwei Wang, Yanjun Cao, et al. Microsaccade-inspired event camera for robotics. *Science Robotics*, 9(90):eadj8124, 2024. 2

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

[13] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 4

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3

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

[16] iniVation. Understanding the performance of neuromorphic event-based vision sensors, 2020. 2

[17] Siddhant Jain, Daniel Watson, Eric Tabellion, Ben Poole, Janne Kontkanen, et al. Video interpolation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7341–7351, 2024. 1, 3

[18] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18032–18042, 2023. 2, 7

[19] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[21] Kazutoshi Kodama, Yusuke Sato, Yuhi Yorikado, Raphael Berner, Kyoji Mizoguchi, Takahiro Miyazaki, Masahiro Tsukamoto, Yoshihisa Matoba, Hirotaka Shinozaki, Atsumi Niwa, et al. 1.22 $\mu$m 35.6 mpixel rgb hybrid event-based vision sensor with 4.88 $\mu$m-pitch event pixels and up to 10k event frame rate by adaptive control on event sparsity. In *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 92–94. IEEE, 2023. 2

[22] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023. 6, 7

[23] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128 $times$128 120 db 15 $mu$ s latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008. 1

[24] 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 *Computer Vision – ECCV 2020*, pages 695–710, Cham, 2020. Springer International Publishing. 3

[25] Chunxu Liu, Guozhen Zhang, Rui Zhao, and Limin Wang. Sparse global matching for video frame interpolation with large motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19125–19134, 2024. 1

[26] Yongrui Ma, Shi Guo, Yutian Chen, Tianfan Xue, and Jinwei Gu. Timelens-xl: Real-time event-based video frame

interpolation with large motion. In *European Conference on Computer Vision*, pages 178–194. Springer, 2024. 2, 7

[27] Yapeng Meng, Taoyi Wang, and Yihan Lin. Technical report of a dmd-based characterization method for vision sensors. *arXiv preprint arXiv:2203.14672*, 2025. 2, 6

[28] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 6, 7

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[30] Basilio Pueo. High speed cameras for motion analysis in sports science. *Journal of Human Sport and Exercise*, 11(1): 53–73, 2016. 1

[31] Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, and Peyman Milanfar. Multiscale structure guided diffusion for image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10721–10733, 2023. 3

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[33] Wei Shang, Dongwei Ren, Yi Yang, Hongzhi Zhang, Kede Ma, and Wangmeng Zuo. Joint video multi-frame interpolation and deblurring under unknown exposure time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13935–13944, 2023. 1

[34] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Blurry video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[35] Xiao Shu and Xiaolin Wu. Real-time high-fidelity compression for extremely high frame rate video cameras. *IEEE Transactions on Computational Imaging*, 4(1):172–180, 2017. 1

[36] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. Xvfi: extreme video frame interpolation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14489–14498, 2021. 6, 7

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 5

[38] Vasanth Subramanyam, Jayendra Kumar, and Shiva Nand Singh. Temporal synchronization framework of machine-vision cameras for high-speed steel surface inspection systems. *Journal of Real-Time Image Processing*, 19(2):445–461, 2022. 1

[39] 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)*, pages 18043–18052, 2023. 2, 3, 7

[40] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16155–16164, 2021. 1, 2

[41] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 5

[42] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024. 3

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[44] Qingyu Wu, Xiaoxiao Li, Kang Wang, and Hazrat Bilal. Regional feature fusion for on-road detection of objects using camera and 3d-lidar in high-speed autonomous vehicles. *Soft Computing*, 27(23):18195–18213, 2023. 1

[45] Yixin Yang, Jinxiu Liang, Bohan Yu, Yan Chen, Jimmy S. Ren, and Boxin Shi. Latency correction for event-guided deblurring and frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24977–24986, 2024. 3

[46] Zheyu Yang, Taoyi Wang, Yihan Lin, Yuguo Chen, Hui Zeng, Jing Pei, Jiazheng Wang, Xue Liu, Yichun Zhou, Jianqiang Zhang, et al. A vision chip with complementary pathways for open-world sensing. *Nature*, 629(8014):1027–1033, 2024. 1, 2, 3, 7

[47] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5682–5692, 2023. 1

[48] 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. 3

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[50] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17765–17774, 2022. 3