SpikeDiff: Zero-shot High-Quality Video Reconstruction from Chromatic Spike Camera and Sub-millisecond Spike Streams

Siqi Yang[1,2,3] Jinxiu Liang[2,3,4*] Zhaojun Huang[2,3] Yeliduosi Xiaokaiti[2,3]
Yakun Chang[5,6] Zhaofei Yu[1,3] Boxin Shi[2,3,1*]

[1] Institute for Artificial Intelligence, Peking University

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

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

[4] National Institute of Informatics [5] Institute of Information Science, Beijing Jiaotong University

[6] Visual Intelligence +X International Cooperation Joint Laboratory of the Ministry of Education

cssherryliang@gmail.com, shiboxin@pku.edu.cn

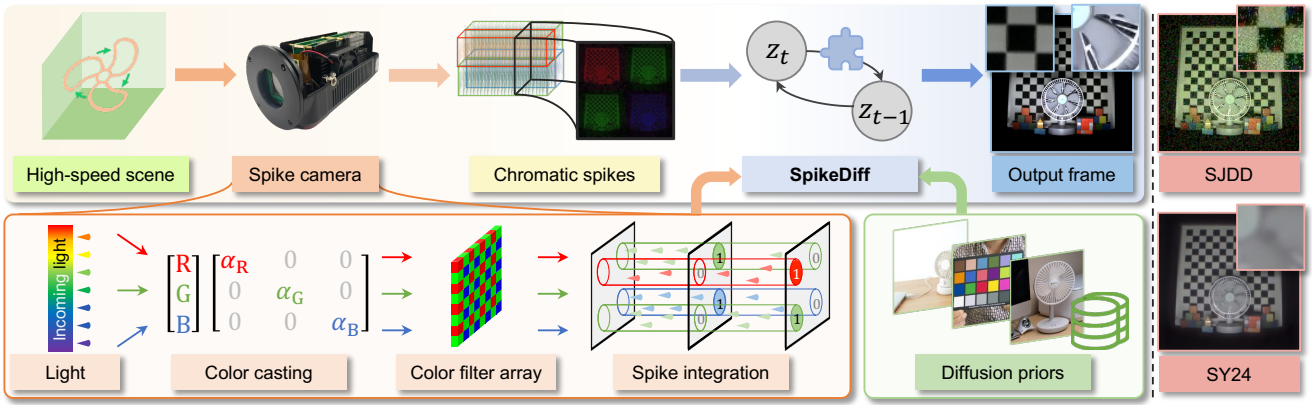


Figure 1. Spike cameras capture chromatic spikes with high temporal resolution, enabling the observation of scenes with rapid motion. As illustrated in the orange region, the spike camera utilizes a color filter array to capture chromatic information and continuously accumulates photons until reaching a threshold that triggers a spike. To reconstruct high-quality videos from these chromatic spikes, we propose **SpikeDiff**, a zero-shot framework that leverages principled priors from pretrained diffusion models and integrates physics-based guidance derived from the spike camera mechanism. SpikeDiff generates visually appealing videos without noise or motion blur, significantly outperforming existing chromatic spike reconstruction methods (*e.g.*, SJDD [11] and CSpkNet [12]).

Abstract

High-speed video reconstruction from neuromorphic spike cameras offers a promising alternative to traditional frame-based imaging, providing superior temporal resolution and dynamic range with reduced power consumption. Nevertheless, reconstructing high-quality colored videos from spikes captured in ultra-short time intervals (sub-millisecond) remain challenging due to the inherently noisy nature of spikes. While some existing methods extend the temporal capture window to improve reconstruction quality, they inevitably compromise the temporal resolution advantages of spike cameras. In this paper, we introduce SpikeDiff, the first zero-shot framework that leverages pretrained diffusion models to reconstruct high-quality colored videos from sub-millisecond (0.5ms) chromatic spike streams. By incorporating physics-based guidance into the diffusion sampling process, SpikeDiff bridges the domain gap between chromatic spikes and conventional images, enabling high-fidelity reconstruction without requiring domain-specific training data. Extensive experiments demonstrate that SpikeDiff achieves impressive reconstruction quality while maintaining ultra-high temporal resolution, outperforming existing methods across diverse challenging scenarios in both perceptual quality and structural preservation.

1. Introduction

High-speed motions–from droplet impacts to mechanical dynamics–occur at speeds surpassing human perceptual processing capabilities. Conventional high-speed cameras are capable of capturing these motions but encounter funda-

* Corresponding authors.

mental trade-offs [37, 45, 51]: higher temporal resolution requires shorter exposures, leading to reduced light capture and limited dynamic range, alongside increased power consumption and larger data volumes. These constraints make continuous high-speed capture over extended periods either impractical or impossible.

Neuromorphic cameras have emerged as a compelling alternative to conventional high-speed imaging, including event cameras [9, 25, 27] and spike cameras [3, 21, 58]. Drawing inspiration from the fovea in biological visual systems, each pixel in a spike camera continuously accumulates photons, generates spikes upon reaching the predefined threshold, and subsequently proceeds to the next accumulation period. Operating at a sampling rate of $20,000$ Hz with 1000×1000 resolution, spike cameras encode visual information with low data overhead and power consumption, compared to conventional cameras with similar sampling rate and resolution. This enables continuously blur-free capture with high-speed motion and high contrast over several minutes that conventional sensors struggle to handle.

For human perception, the binary representations of spikes are inherently difficult to interpret and require reconstruction into video frames [54, 55, 58]. However, reconstructing high-quality video frames from spikes remains challenging, particularly within ultra-short time intervals where limited spikes make existing methods highly susceptible to *noise* [57]. This challenge intensifies for color reconstruction, as Color Filter Arrays (CFAs) further reduce the number of available spikes within short time intervals.

Existing approaches attempt to mitigate these challenges by extending the observed time intervals to accumulate more spikes [31, 56, 57]. This strategy inadvertently introduces *motion blur* and *ghosting artifacts* that compromise the temporal resolution advantage of spike cameras. While recent deep learning-based methods have shown promise in optimizing this trade-off problem, they typically exhibit limited generalization capabilities [10, 11], require extensive training data [12], or demand longer time intervals [48]. Yet, these methods still struggle to completely eliminate the noise [11] (upper right in Figure 1) or avoid introducing motion blur in high-speed motion regions [48] (lower right in Figure 1).

The recent success of diffusion models in image and video restoration [15, 22, 43, 49] suggests a promising direction. These methods have demonstrated remarkable capability in encoding rich priors about natural scene statistics and temporal dynamics. This raises an intriguing question: *Can we leverage these powerful priors to reconstruct high-quality frames and videos from extremely noisy chromatic spikes captured within sub-millisecond* [1] *time intervals?*

Integrating pretrained diffusion models with spike-to-video reconstruction presents fundamental challenges: Spike cameras capture visual information through asynchronous voltage accumulation and threshold-based sampling—a process fundamentally different from the fixed-interval capturing mechanism of frame-based cameras. Furthermore, chromatic spikes represent raw sensor measurements without established signal processing pipelines, while pretrained diffusion models are typically trained on processed sRGB frames with optimized color and tone characteristics. These disparities create a significant domain gap that must be addressed for effective reconstruction.

In this paper, we propose **SpikeDiff**, the *first* zero-shot spike-to-video reconstruction framework that leverages pretrained diffusion models as principled *priors* for high-quality colored video reconstruction from chromatic spikes. By establishing a physically grounded likelihood model of spike measurements throughout the reverse diffusion process, SpikeDiff approximates the posterior sampling of high-quality videos from sub-millisecond spike streams. Unlike existing approaches that require extensive domain-specific training data and risk overfitting, our method operates in a zero-shot manner, eliminating both the computational burden of training and the need for large-scale datasets.

Our key contributions include:

- the first framework that leverages pretrained diffusion models as powerful priors for high-quality, zero-shot video reconstruction from chromatic spikes in ultra-short intervals, eliminating the need for domain-specific training data; and
- a physics-based and differentiable formulation of the spike generation process, which enables effective integration with pretrained diffusion models, bridging the domain gap between spikes and conventional images.

Extensive experiments across diverse challenging scenarios demonstrate that SpikeDiff consistently outperforms prior methods that rely on extended time intervals and compromising temporal resolution, effectively breaking the traditional quality-time resolution trade-off in spike-to-video reconstruction (see our result shown in Figure 1).

2. Related work

Video reconstruction with monochromatic spikes. Reconstructing high-speed videos from spikes typically employs the imaging model of spike cameras [21]. By accumulating the spikes triggered within a certain time window or estimating the interval between successive spikes, textures can be reconstructed [57, 58]. However, due to the limited number of photons arriving at the pixel in an ultra-short time, such straightforward reconstruction suffers from significant noise. To address this issue, Chang *et al*. [4] propose enhancing the signal-to-noise ratio (SNR) by combining multi-bit spikes with binary spikes in a rolling-mixed-bit manner. However, this solution requires hardware modifications, limiting its practicality in more general scenes. Zhao *et al*. [54] present

[1] Existing state-of-the-art methods [10–12] require at least a time interval of $40/20,000$ Hz $= 2$ ms for high-quality video reconstruction, while we aim at ultra-short time interval of $10/20,000$ Hz $= 0.5$ ms.

Spk2ImgNet, a hierarchical architecture that progressively fuses the spikes, offering an alternative approach to improve reconstruction quality. However, its generalization ability is limited since it relies on synthetic data for training. To address the issue of real-world ground truth for training, self-supervised methods [5, 6] are developed to reduce dependence on synthetic datasets. However, it is still challenging to reconstruct color videos from monochromatic spikes. From the perspective of incorporating additional color guidance, Chang *et al*. [3] develop a hybrid spike-RGB camera system that performs spatial alignment and frame interpolation simultaneously, enabling the recovery of 1000 FPS color video. However, this hybrid camera system necessitates synchronization and optical alignment, while the beam splitter also poses challenges for constructing a compact device.

Demosaicking and denoising for unconventional cameras. Various methods for joint demosaicking and denoising have been proposed to obtain high-resolution and low-noise color images from frame-based RAW data, ranging from traditional model-based techniques [1, 8, 19, 23, 32–34, 36, 50] to more recent data-driven approaches [18, 24, 29, 41, 42, 46, 53]. However, all these methods cannot be applied to binary data captured by the new-generation high-speed cameras, such as event cameras [9, 25, 27], quanta image sensors [16, 17], and spike cameras [21]. In the case of event cameras, Xu *et al*. [47] and Lu *et al*. [30] present Transformer-based architectures for demosaicking missing pixel values in RAW domain processing. However, due to the sparsity of events, these approaches still rely on frame-based cameras for texture reconstruction. In the context of quanta image sensors, Elgendy *et al*. [14] integrate a frequency selection method with a deep neural network based filtering approach, leveraging the luma channels to assist in the denoising of the chroma channels. Ma *et al*. [31] design a blue-noise pseudo-random RGBW color filter array and successfully reconstruct high-quality color images from mosaicked single-bit frames, even in high-dynamic-range (HDR) scenes with complex and rapid motion. For spike cameras, Dong *et al*. [11] propose an offset-sharing deformable convolution module to align temporal features of color channels and develop a spike noise estimator to characterize the noise distribution. However, this method is trained on synthetic data, limiting its generalization to real-world data. Yang *et al*. [48] introduce a self-supervised denoising module trained exclusively on real-world chromatic spikes, achieving 2000 FPS color video reconstruction. However, there remains significant research potential in balancing noise suppression with detail preservation.

3. Method

We present the degradation model of chromatic spikes based on the physical formulation and propose SpikeDiff, a zero-shot framework for reconstructing consecutive high-quality frames from chromatic spikes. This approach leverages pretrained diffusion models as principled priors to effectively address the inherent ill-posedness of the problem.

3.1. Preliminaries

Chromatic spikes. The chromatic spike camera leverages a spike sensor equipped with a Bayer-pattern (RGGB) CFA to capture the scene with chromatic information. As shown in Figure 1, when photons filtered by the CFA reach a pixel, the electrons generated by these photons are continuously accumulated as long as the voltage does not reach the threshold E_{th}. Simultaneously, the readout circuit samples the pixel values at a frequency of $20,000$ Hz and a signal of 0 is read out at each readout point. Once the accumulated voltage reaches E_{th}, a signal of 1 is read out and the voltage of this pixel is reset to continue the next accumulation period. For each pixel i, we denote the accumulated voltage at a readout point τ as $E(i,\tau)$, the triggered spike at τ is:

$$S(i,\tau) = \begin{cases} 1, & \text{if } E(i,\tau) \geq E_{\text{th}}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Diffusion models. Latent diffusion models (LDMs) operate in a compressed latent space learned by an autoencoder pair $(\mathcal{E}, \mathcal{D})$, where $\boldsymbol{X} = \mathcal{D}(\boldsymbol{Z})$ and $\boldsymbol{Z} = \mathcal{E}(\boldsymbol{X})$. An LDM learns to reverse a diffusion process that gradually adds Gaussian noise to a latent vector $\boldsymbol{Z}_0$. The reverse process, modeled by a stochastic differential equation (SDE), generates a clean latent $\boldsymbol{Z}_0$ from pure noise $\boldsymbol{Z}_T \sim \mathcal{N}(0, I)$ [40]:

$$d\boldsymbol{Z} = [-f(\boldsymbol{Z}, t) - g^2(t)\nabla_{\boldsymbol{Z}_t} \log p_t(\boldsymbol{Z}_t)]dt + g(t)dw, \tag{2}$$

where $f(\cdot, t)$ and $g(t)$ are drift and diffusion coefficients, w represents the standard Brownian motion. The key component is the score function $\nabla_{\boldsymbol{Z}_t} \log p_t(\boldsymbol{Z}_t)$, which is approximated by a time-conditioned neural network $\epsilon_\theta(\boldsymbol{Z}, t)$. By iteratively applying the learned score, the reverse SDE can sample from the learned data prior $p(\boldsymbol{Z}_0)$ to generate a high-quality image $\boldsymbol{X}_0 = \mathcal{D}(\boldsymbol{Z}_0)$.

3.2. Diffusion-based reconstruction from spikes

Reconstructing consecutive high-quality frames from real-captured chromatic spike streams is challenging due to the non-negligible noise perturbations, particularly within extremely limited time intervals (*e.g*., sub-millisecond). To address this challenge, we leverage recent advancements in diffusion models, which incorporate learned priors regarding the distribution of high-quality images, and formulate the reconstruction of video frames from chromatic spikes as a Maximum a Posteriori (MAP) estimation problem:

$$\begin{aligned} \boldsymbol{X}^* &= \text{argmax}_{\boldsymbol{X}}\, p(\boldsymbol{X}|\boldsymbol{Y}) \\ &= \text{argmax}_{\boldsymbol{X}}\, p(\boldsymbol{Y}|\boldsymbol{X})p(\boldsymbol{X}), \end{aligned} \tag{3}$$

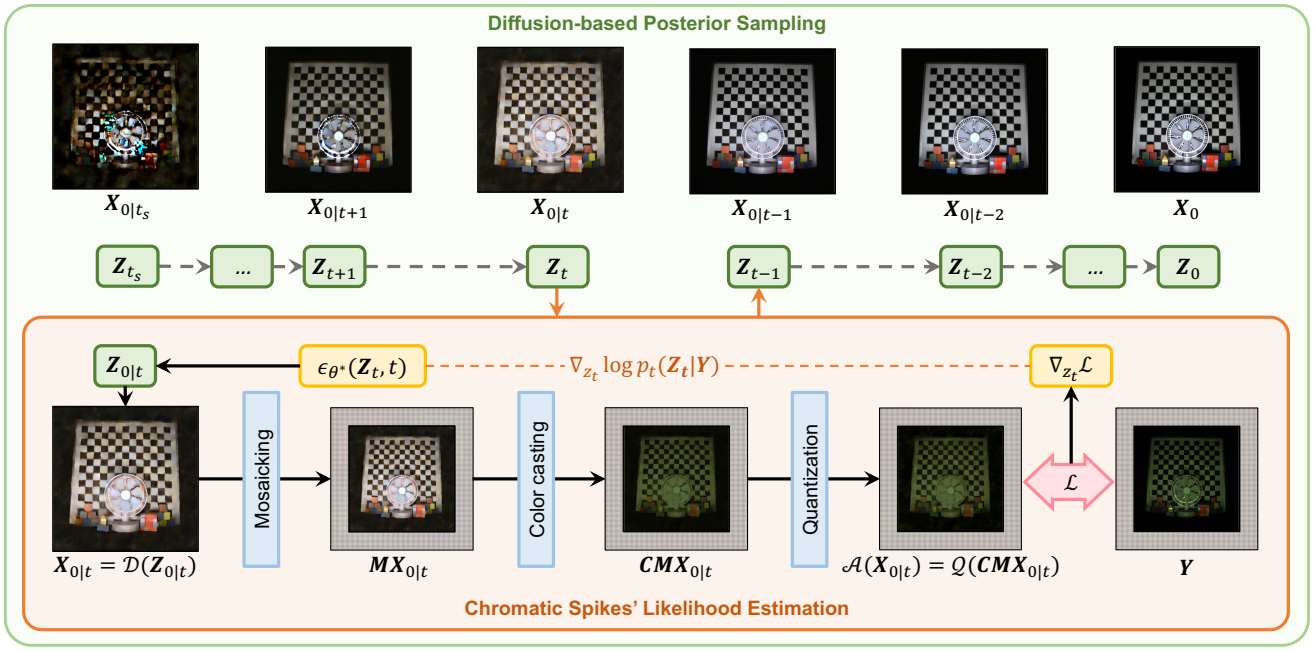


Figure 2. Pipeline of SpikeDiff. Our diffusion-based posterior sampling procedure follows the iterative denoising process of diffusion models, starting from the initialization $\boldsymbol{Z}_{t_s}$ and gradually approaching the desired latent $\boldsymbol{Z}_0$, indicated by the dashed lines. To maximize the likelihood of the reconstructed video given the spike observations $\boldsymbol{Y}$, we enhance each iteration with physics-based spike guidance, which comprises three specially-designed differentiable degradation operators: mosaicking $\boldsymbol{M}$, color casting $\boldsymbol{C}$, and quantization $\mathcal{Q}$. Finally, a multiscale loss function bridges the gap between degraded frames $\mathcal{A}(\boldsymbol{X}_{0|t})$ and observed SFR frames $\boldsymbol{Y}$. As shown in the top sequence, the visual quality of generated frames progressively improves until convergence. Note that the internal degraded frames and the SFR frames are all Bayer-patterned mosaic images; we present their corresponding demosaicked results for better visualization.

where the prior $p(\boldsymbol{X})$ encodes knowledge of natural images, and the likelihood $p(\boldsymbol{Y}|\boldsymbol{X})$ represents the data fidelity to ensure the reconstruction is faithful to the observed chromatic spikes $\boldsymbol{Y}$.

The degradation model, mapping light intensity to spike signals (illustrated in Figure 1), establishes the connection from RGB frames $\boldsymbol{X}$—the modality employed in pretrained diffusion models—and chromatic spikes $\boldsymbol{Y}$, and thereby deriving likelihood $p(\boldsymbol{Y}|\boldsymbol{X})$ illustrated in the next (Sec. 3.3). This establishes a principled bridge that enables the reconstruction of $\boldsymbol{X}$ from $\boldsymbol{Y}$ via posterior sampling in the MAP framework, as illustrated in the green box in Figure 2, where diffusion models serve as learned priors for $\boldsymbol{X}$ while the likelihood ensures fidelity to $\boldsymbol{Y}$. This is exactly the key in our method to connect the modality used in pretrained diffusion models, RGB frames, and chromatic spikes.

Given the significant disparity between chromatic spike streams and conventional images, the data fidelity term $p(\boldsymbol{Y}|\boldsymbol{X})$ cannot be directly derived from time-continuous chromatic spikes. Instead, we resort to a more compact representation. According to the integrating and firing mechanism of spikes, the light intensity is proportional to the spike firing rate (SFR) [57]. Without considering noise, the SFR at τ' can be counted through collecting a number of spikes within a predefined time interval of size $\Delta\tau$:

$$Y(i,\tau') = \frac{1}{\Delta\tau} \cdot \sum_{\tau=\tau'-\Delta\tau/2}^{\tau'+\Delta\tau/2} S(i,\tau). \tag{4}$$

By collecting the SFRs from the pixel array at τ', we obtain mosaicked SFR frames $\boldsymbol{Y} \in \mathbb{R}^{T\times H\times W}$, where T denotes the number of frames, and H, W represent the height and width, respectively. The SFR frames, estimated from real-captured chromatic spikes, are sensitive to the inherent noise, especially within limited time intervals, and exhibit misalignment with corresponding consecutive high-quality frames. Consequently, we propose a corresponding degradation model, comprising differentiable operators, as detailed in Sec. 3.3, to bridge the gap between $\boldsymbol{X}$ and $\boldsymbol{Y}$ and resolve the likelihood estimation required for posterior sampling, as illustrated in the orange box in Figure 2.

Posterior sampling. The primary challenge in reconstructing consecutive high-quality frames from sub-millisecond chromatic spikes lies in its inherent ill-posedness, making the prior term $p(\boldsymbol{X})$ crucial. We leverage diffusion models as powerful priors. Given the prior distribution of natural images $p(\boldsymbol{Z}_t)$ encoded by the pretrained latent diffusion model in Eq. (2) and likelihood defined in the next, we solve the MAP problem in Eq. (3) by sampling from the posterior

distribution $p(\boldsymbol{X}|\boldsymbol{Y})$. This is achieved by modifying the reverse SDE of the diffusion model to incorporate guidance from our observations. Using Bayes' rule, the score of the posterior distribution $p_t(\boldsymbol{Z}_t|\boldsymbol{Y})$ can be written as:

$$\nabla_{\boldsymbol{Z}_t} \log p_t(\boldsymbol{Z}_t|\boldsymbol{Y}) = \nabla_{\boldsymbol{Z}_t} \log p(\boldsymbol{Z}_t) + \nabla_{\boldsymbol{Z}_t} \log p(\boldsymbol{Y}|\boldsymbol{Z}_t), \tag{5}$$

where the prior term $\nabla_{\boldsymbol{Z}_t} \log p(\boldsymbol{Z}_t)$ is given by the pre-trained network $\epsilon_{\theta^*}(\boldsymbol{Z}_t, t)$. The likelihood term $p(\boldsymbol{Y}|\boldsymbol{Z}_t)$ guides the sampling towards latents $\boldsymbol{Z}_t$ that are likely to produce our observed SFR frame $\boldsymbol{Y}$. The full reverse SDE for posterior sampling is:

$$\begin{aligned} d\boldsymbol{Z} = [&- f(\boldsymbol{Z}, t) - g^2(t)(\nabla_{\boldsymbol{Z}_t} \log p(\boldsymbol{Z}_t) \\ &+ \nabla_{\boldsymbol{Z}_t} \log p(\boldsymbol{Y}|\boldsymbol{Z}_t))]dt + g(t)dw. \end{aligned} \tag{6}$$

To compute the likelihood score, we must relate the noisy latent $\boldsymbol{Z}_t$ at an intermediate timestep t to the observation $\boldsymbol{Y}$. This involves approximating the conditional probability $p(\boldsymbol{Y}|\boldsymbol{Z}_t)$, which can be expressed as:

$$\begin{aligned} p(\boldsymbol{Y}|\boldsymbol{Z}_t) &= \int p(\boldsymbol{Y}|\boldsymbol{Z}_0, \boldsymbol{Z}_t)\, p(\boldsymbol{Z}_0|\boldsymbol{Z}_t)\, d\boldsymbol{Z}_0 \\ &= \int p(\boldsymbol{Y}|\boldsymbol{Z}_0)\, p(\boldsymbol{Z}_0|\boldsymbol{Z}_t)\, d\boldsymbol{Z}_0. \end{aligned} \tag{7}$$

This integral is intractable. Instead, we approximate it by first estimating the expected clean latent $\boldsymbol{Z}_{0|t} = \mathbb{E}[\boldsymbol{Z}_0|\boldsymbol{Z}_t]$ from $\boldsymbol{Z}_t$. For diffusion models like DDPM [20], the forward process is

$$\boldsymbol{Z}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{Z}_0 + \sqrt{1-\bar{\alpha}_t}z, \quad z \sim \mathcal{N}(0, I), \tag{8}$$

where $\bar{\alpha}_t$ is the noise schedule parameter from the diffusion model (*e.g.*, DDPM [20]). Using Tweedie's formula [7], we can estimate the conditional mean:

$$\begin{aligned} \boldsymbol{Z}_{0|t} &= \mathbb{E}[\boldsymbol{Z}_0|\boldsymbol{Z}_t] \\ &= \frac{1}{\sqrt{\bar{\alpha}_t}}(\boldsymbol{Z}_t + (1-\bar{\alpha}_t)\nabla_{\boldsymbol{Z}_t} \log p_t(\boldsymbol{Z}_t)). \end{aligned} \tag{9}$$

Here, we replace the true score $\nabla_{\boldsymbol{Z}_t} \log p_t(\boldsymbol{Z}_t)$ with its neural network approximation $\epsilon_{\theta^*}(\boldsymbol{Z}_t, t)$. With this estimate of the clean latent, $\boldsymbol{Z}_{0|t}$, following the expression $p(\boldsymbol{Y}|\boldsymbol{Z}_t) = \mathbb{E}_{\boldsymbol{Z}_0 \sim p(\boldsymbol{Z}_0|\boldsymbol{Z}_t)}[p(\boldsymbol{Y}|\mathcal{D}(\boldsymbol{Z}_0))]$, we can use a surrogate likelihood approach [7] to make the likelihood tractable:

$$p(\boldsymbol{Y}|\boldsymbol{Z}_t) \approx p(\boldsymbol{Y}|\mathcal{D}(\mathbb{E}[\boldsymbol{Z}_0|\boldsymbol{Z}_t])). \tag{10}$$

This formulation provides a practical way to compute the likelihood gradient.

3.3. Chromatic spikes' likelihood estimation

Given a set of sequential SFR frames, denoted by $\boldsymbol{Y}$, our objective is to establish a connection between the SFR frames and the desired target reconstructions $\boldsymbol{X}$, with both $\boldsymbol{Y}$ and $\boldsymbol{X}$ normalized to the range $[0, 1]$. Hence, we resolve the likelihood estimation of $p(\boldsymbol{Y}|\boldsymbol{Z}_t)$ in Eq. (10) and thereby complete the diffusion-based posterior sampling strategy.

Degradation model of chromatic spikes. The physical process generating SFR frames $\boldsymbol{Y}$ can be modeled as a non-linear inverse problem: $\boldsymbol{Y} = \mathcal{A}(\boldsymbol{X}) + \boldsymbol{N}$, where $\mathcal{A}(\cdot)$ denotes the known forward operator modeling the degradation process, and $\boldsymbol{N}$ represents stochastic noise. This problem is inherently ill-posed due to irreversible information loss during the imaging process and interference from noise in the capture system.

The forward operator $\mathcal{A}(\cdot)$ in the imaging pipeline transforms the latent high-quality frame $\boldsymbol{X}$ through a sequence of operations: $\mathcal{A}(\boldsymbol{X}) = \mathcal{Q}(\boldsymbol{C}\boldsymbol{M}\boldsymbol{X})$, where $\mathcal{Q}(\cdot)$ represents pixel-wise quantization, $\boldsymbol{C}$ implements pixel-wise color casting with per-channel scaling, and $\boldsymbol{M} : \mathbb{R}^{T\times H\times W\times 3} \rightarrow \mathbb{R}^{T\times H\times W}$ denotes the mosaicking operator, modeling the color filter array pattern. This decomposition explicitly models each step of the image formation process, providing a foundation for our reconstruction approach.

Under the assumption of Gaussian noise, we model the likelihood $p(\boldsymbol{Y}|\mathcal{D}(\boldsymbol{Z}_{0|t}))$ of observing SFR frames $\boldsymbol{Y}$ given the target frame $\boldsymbol{X}$ as a Gaussian distribution. Maximizing this likelihood translates to minimizing the following loss:

$$\mathcal{L}(\boldsymbol{Z}_t) = \frac{1}{\sigma^2}\|\boldsymbol{Y} - \mathcal{A}(\mathcal{D}(\boldsymbol{Z}_{0|t}))\|_2^2, \tag{11}$$

where σ represents the standard deviation of the noise $\boldsymbol{N}$, and $\mathcal{A}$ represents the differentiable degradation operators defined later. By substituting the score function into Eq. (5), we obtain:

$$\nabla_{\boldsymbol{Z}_t} \log p_t(\boldsymbol{Z}_t|\boldsymbol{Y}) = \epsilon_{\theta^*}(\boldsymbol{Z}_t, t) - s\nabla_{\boldsymbol{Z}_t}\mathcal{L}(\boldsymbol{Z}_t), \tag{12}$$

where s controls the strength of measurement guidance.

Differentiable degradation operators. A key innovation in SpikeDiff is our design of differentiable operators in $\mathcal{A}$ that model the physical characteristics of chromatic spike cameras, which include color casting $\boldsymbol{C}$, quantization $\mathcal{Q}$, and mosaicking $\boldsymbol{M}$. Based on the gray-world prior [2] for the target frame, which indicates that each color channel should have a mean value close to one another, we estimate the color casting coefficients $\alpha_R, \alpha_G, \alpha_B$ as $\alpha_c = \frac{1}{\mu(\boldsymbol{Y}_c)}$, where $c \in \{\text{R,G,B}\}$ denotes the color channel, $\mu(\boldsymbol{Y}_c)$ is the mean value of color channel c among $\boldsymbol{Y}$ across the temporal axis. Then the channel-wise scaling factors for the color casting operator $\boldsymbol{C}$ can be computed as η/α_c, where parameter η is used to adjust the value range.

SFR frames are inherently quantized and discrete as they are accumulation of binary spike frames, which differ from

8-bit frames $\boldsymbol{X}$. To simulate the quantization effect and maintain the backward differentiability, we design a soft quantization operator $\mathcal{Q}$. Given the expected quantization levels determined by SFR estimation, the soft quantization function $q(\cdot)$ operates on each pixel individually:

$$q(x) = l(x) \cdot (1 - \varphi(x)) + u(x) \cdot \varphi(x), \tag{13}$$

$$\varphi(x) = [\tanh(k(x - \frac{l(x) + u(x)}{2})) + 1]/2, \tag{14}$$

where $l(x)$ and $u(x)$ denote the lower and upper quantization levels of x, $\varphi(\cdot)$ quantizes the input value to the range $[0, 1]$ with a smooth curve, and hyperparameter k controls the smoothness, as shown in Figure 3(a).

The naive Bayer-pattern mosaicking operator $\boldsymbol{M}$ converts target frame from dimension $T \times H \times W \times 3$ to $T \times \frac{H}{2} \times \frac{W}{2} \times 4$, resulting in a partially defined backpropagation function due to its gradient discontinuity. To simulate the Bayer-pattern CFA while maintaining a dense gradient flow, we extend the operator's derivative with linear interpolation convolution, similar to demosaicking algorithms.

By composing these operators, we conclude a differentiable simulation from consecutive frames to SFR frames, enabling optimization in the posterior sampling process.

Multiscale enhancement. To preserve fine spatial details, we incorporate a Laplacian pyramid $\mathcal{P}(\boldsymbol{X})$ that guides the diffusion process across multiple scales. The Laplacian pyramid decomposes the generated consecutive frames into a multiscale representation of frequency components, facilitating the preservation of fine details critical for high-quality output. The Laplacian pyramid is calculated by subtracting the upsampled version of the downsampled frames from the original frames:

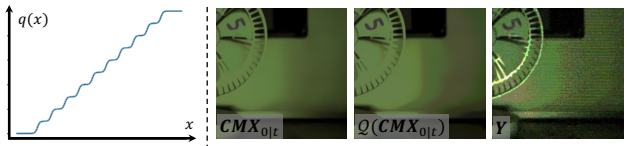
$$\mathcal{P}(\boldsymbol{X}) = \{\boldsymbol{X}_l - \mathbf{U}(\mathbf{D}(\boldsymbol{X}_l))\}_{l=1}^{L}, \ \boldsymbol{X}_l = \mathbf{U}(\mathbf{D}(\boldsymbol{X}_{l-1})) \tag{15}$$

where $\mathbf{D}$ and $\mathbf{U}$ are the upsampling operator and downsampling operator, respectively, and L is the number of pyramid levels. This multiscale decomposition effectively guides the diffusion process across different frequency bands, ensuring preservation of high-frequency details in reconstruction.

4. Experiments

4.1. Implementation details

Quantization levels. As introduced in Section 3.2, SFR frames are accumulations over predefined time intervals $\Delta\tau$. Given that the spikes are binary, the value range of SFR frames is restricted to $\{\frac{u}{\Delta\tau} | u \in \mathbb{N} \cap [0, \Delta\tau]\}$. For instance, we set $\Delta\tau = 10$ in our experiments to achieve sub-millisecond reconstruction, and the corresponding quantization levels are $0, 0.1, 0.2, \ldots, 0.9, 1$, leading to the soft quantization function visualized in Figure 3(a).



(a) Visualization of $q(x)$ (b) Quantization effectiveness

Figure 3. Visualization of the quantization operator. (a) The soft quantization function $q(x)$ for SFR frames with $\Delta\tau = 10$. (b) Detailed demonstration of quantization effectiveness on continuous pixel values.

Color casting estimation. To obtain better color casting coefficients and maintain consistency across video frames reconstruction, we average the estimations of α_c from all SFR frames in the sequence, and apply the same α_c throughout the video reconstruction.

Pretrained diffusion model. We implement SpikeDiff upon the pretrained weights from [28], which is adapted from Stable Diffusion v2.1 [38], selected for their strong priors and close alignment with our task. We extend the image-to-image pipeline to spikes-to-video by dividing SFR frames into small chunks, and achieve temporal consistency by leveraging high-fidelity reconstruction from time-continuous chromatic spikes with physics-based guidance.

4.2. Evaluation dataset

As our proposed method leverages pretrained diffusion models, which contain principled priors of the real world images and videos, we qualitatively evaluate the performance of our method on a real-world dataset. To this end, we collect a series of real chromatic spikes within the spike camera. The spike camera is equipped with a Bayer-pattern color filter array and a 1000×1000 pixel spike sensor. Limited by the generation ability of pretrained diffusion models and resolution of compared methods [48], we spatially crop the chromatic spikes into a lower resolution of 512×512, maintaining the same Bayer-pattern (RGGB). The collected dataset contains chromatic spikes from various scenes with different levels of motion and color variation. We would like to contribute this dataset to the community for further research.

Despite the real-captured chromatic spikes, we also employ the chromatic spikes simulator described in CSpkNet [12] and SY24 [48] to generate synthetic chromatic spikes from real-world high-frame-rate videos. We collect over 100 dynamic scenarios from X4K1000FPS [39] dataset, and generate the corresponding chromatic spikes with the simulator. The synthetic dataset is used to conduct the quantitative evaluation of our proposed method, which further demonstrates the superiority of SpikeDiff.

4.3. Qualitative evaluation

To demonstrate the superiority of our proposed method, we conduct the qualitative evaluation on chromatic spikes captured from the real world, comparing SpikeDiff with exist-

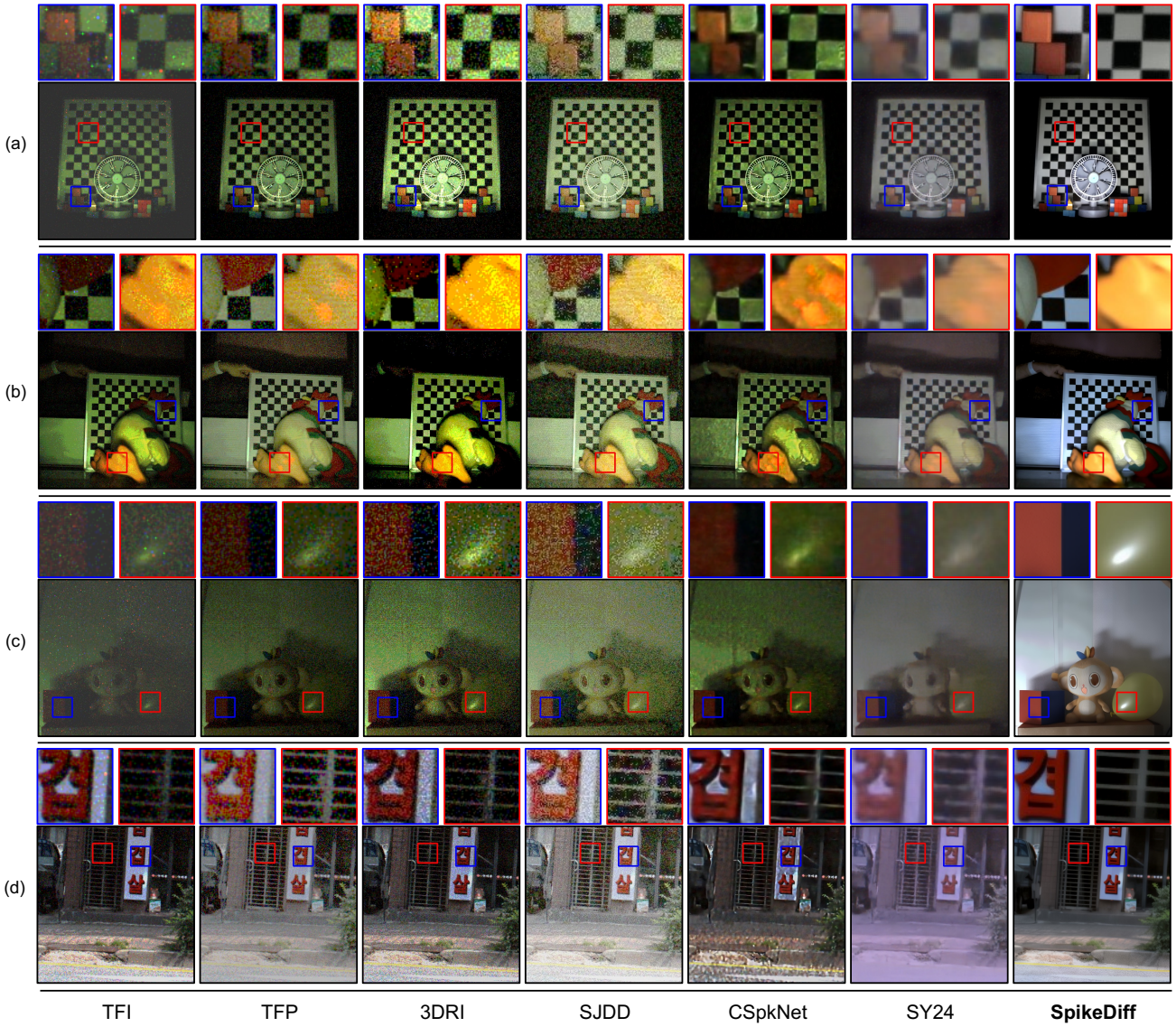


Figure 4. Qualitative comparison on real-world captured chromatic spikes (a-c), and synthetic chromatic spikes simulated from real-world high-frame-rate videos (d). Compared methods include TFI [57], TFP [57], 3DRI [10], SJDD [11], CSpkNet [12], and SY24 [48]. Details in red / blue bounding boxes are shown on the top. All results are reconstructed from sub-millisecond spike streams (0.5ms), where most methods suffer from severe noise or extreme degradation while SpikeDiff recovers clean and visually pleasing results.

ing chromatic spike reconstruction methods. Our proposed method can recover high-quality color frames from sub-millisecond time intervals (10 spike frames, corresponding to 0.5 ms), which none of the existing neural network based method achieve. We adapt and evaluate the existing chromatic spike reconstruction methods (3DRI [10], SJDD [11], CSpkNet [12], and SY24 [48]), combining the SFR estimation methods (TFI [57], TFP [57]) with demosaicking as baselines. We conduct all experiments on sub-millisecond time intervals (0.5ms) to maintain the fairness in comparison. As shown in Figure 4(a), where a fan is fast rotating in front of the checkerboard, SpikeDiff produces the most clean and visually pleasant results. In contrast, other methods, except CSpkNet and SY24, contain significantly noisy pixels as highlighted in the red and blue bounding boxes. The results of CSpkNet and SY24 contain blurry or degraded artifacts. As the time intervals are strictly limited, all methods are free of motion blur naturally. As for Figure 4(b), we capture a static scene in front of the lamp with the spike camera slightly shaking, to study the reconstruction quality under hard light. Compared to others, SpikeDiff generates the most clear and sharp details in blue bounding box. And SpikeDiff produces a smoother texture under the direct illumination in the red bounding box, while most of other methods show obvious quantized artifacts. In Figure 4(c), we leverage the spike streams captured by Yang *et al.* [48], where SpikeDiff outperforms all other methods in both noise suppression and visual quality. We also illustrate a synthetic scene in Figure 4(d), which further demonstrates the superior reconstruction capability of SpikeDiff on chromatic spikes

Table 1. Quantitative comparison of the proposed method with existing chromatic spike reconstruction methods. The best and second-best results are highlighted in **red** and blue, respectively.

Method	PSNR↑	SSIM↑	FID↓	NIQE↓	IL-NIQE↓
SpikeDiff (Ours)	**18.694**	**0.750**	**2.880**	**5.173**	**38.535**
SY24 [48]	13.013	0.536	21.472	11.592	78.672
SJDD [11]	11.358	0.325	26.009	11.278	44.711
3DRI [10]	18.512	0.390	16.036	8.541	41.816
CSpkNet [12]	13.364	0.641	4.699	5.819	39.948
TFP [57]	12.909	0.559	3.256	12.001	64.554
TFI [57]	14.640	0.552	7.558	12.518	55.758

Table 2. Temporal consistency evaluation, reporting Frame CLIP Score (Frame C.S.), Interpolation Error (Inp. Err.), and Interpolation PSNR (Inp. PSNR) following the metrics from [26]. The best and second-best are highlighted in **red** and blue, respectively.

Method	Frame C.S.↑	Inp. Err.↓	Inp. PSNR↑
SpikeDiff (Ours)	**0.969**	0.057	25.45
SY24 [48]	0.963	**0.041**	**28.65**
SJDD [11]	0.962	0.113	19.34
3DRI [10]	0.965	0.104	19.87
CSpkNet [12]	0.966	0.059	25.27

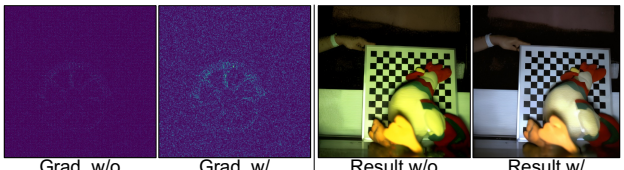


Figure 5. Analysis of differentiable degradation operators. Generated frames receive sparse gradient from chromatic spikes without mosaicking operator. Similarly, reconstructions exhibit incorrect color distribution without the color casting operator.

simulated from real-world videos. Note that post-processing (*e.g.*, color casting) cannot eliminate noise or artifacts from other methods' results. Please refer to our supplementary material for more comparison and detailed discussion.

4.4. Quantitative evaluation

As shown in Table 1, we quantitatively evaluate the performance of SpikeDiff and existing chromatic spike reconstruction methods on the synthetic dataset. Similar to our qualitative evaluation, we evaluate these methods on sub-millisecond time intervals (0.5ms). Our method achieves the best performance across all the metrics, including PSNR, SSIM [44], FID [13], NIQE [35], and IL-NIQE [52], which demonstrates the superiority of SpikeDiff, especially in terms of perceptual metrics that reflect visual quality. As shown in Table 2, we also evaluate the temporal consistency metrics of SpikeDiff and existing reconstruction methods, where SpikeDiff achieves the best or comparable performance.

4.5. Ablation study

To investigate the effectiveness of our proposed degradation operators, we visualize the internal results after applying each operator. The mosaicking operator $\boldsymbol{M}$ converts the generated frames into Bayer-pattern images, which is the beginning step. As shown in Figure 5, without the specific backward formulation for $\boldsymbol{M}$, the gradient map is not well-defined. The color casting operator $\boldsymbol{C}$ is responsible for color correction, which degrades the normal colored image to a greener one (which is more close to the SFR frames) as shown in Figure 5. For the soft quantization operator $\mathcal{Q}$, its effectiveness is demonstrated in Figure 3(b), quantizing the continuous white pixels into discrete levels, similar to the pattern of the SFR frames. As shown in Table 3, we further demonstrate the effectiveness of each degradation operator via quantitative evaluation.

Table 3. Quantitative ablation study on the contribution of different degradation operators to reconstruction quality.

Method	PSNR↑	SSIM↑	FID↓	NIQE↓	IL-NIQE↓
SpikeDiff	18.694	0.750	2.880	5.173	38.535
w/o $\boldsymbol{C}$	16.752	0.729	12.064	6.072	44.661
w/o $\boldsymbol{M}$	17.481	0.723	4.392	6.203	41.356
w/o $\mathcal{Q}$	17.460	0.717	3.938	6.318	40.487

5. Conclusion

We propose the first zero-shot chromatic spike reconstruction method, named SpikeDiff, to recover consecutive clean and high-quality colored frames from sub-millisecond spikes. SpikeDiff leverages the deep priors from pretrained diffusion models to address the extreme noise under very limited time intervals and generate visually pleasant results. SpikeDiff integrates the physics-based chromatic spikes' likelihood into the diffusion-based posterior sampling process, which bridges the domain gap between spikes and conventional images. We conduct experiments on both synthetic and real-captured data to demonstrate the superior performance of SpikeDiff over existing reconstruction methods.

Limitations and future work. As shown in Figure 4(b), there exist "overexposure" cases in the central region under extreme direct illumination. This imperfection is due to the predefined spike firing threshold, resulting in consistent spikes among these pixels. We observed that it is hard to generate ideal details without any explicit guidance. In future research, we plan to explore more degradation model designs and diffusion techniques to handle this issue and further extend the HDR capability of chromatic spike reconstruction.

Acknowledgement

This work is supported by National Natural Science Foundation of China (Grant No. 62302019, 62088102, 62136001), Beijing Natural Science Foundation (Grant No. L233024), and Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012). PKU-affiliated authors thank openbayes.com for providing computing resources.

References

[1] Antoni Buades, Bartomeu Coll, Jean-Michel Morel, and Catalina Sbert. Self-similarity driven color demosaicking. *IEEE TIP*, 18(6):1192–1202, 2009. 3

[2] Gershon Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1): 1–26, 1980. 5

[3] Yakun Chang, Chu Zhou, Yuchen Hong, Liwen Hu, Chao Xu, Tiejun Huang, and Boxin Shi. 1000 FPS HDR video with a Spike-RGB hybrid camera. In *CVPR*, pages 22180–22190, 2023. 2, 3

[4] Yakun Chang, Yeliduosi Xiaokaiti, Yujia Liu, Bin Fan, Zhaojun Huang, Tiejun Huang, and Boxin Shi. Towards HDR and HFR video from rolling-mixed-bit spikings. In *CVPR*, pages 25117–25127, 2024. 2

[5] Shiyan Chen, Chaoteng Duan, Zhaofei Yu, Ruiqin Xiong, and Tiejun Huang. Self-supervised mutual learning for dynamic scene reconstruction of spiking camera. In *IJCAI*, page 2859–2866, 2022. 3

[6] Shiyan Chen, Zhaofei Yu, and Tiejun Huang. Self-supervised joint dynamic scene reconstruction and optical flow estimation for spiking camera. In *AAAI*, pages 350–358, 2023. 3

[7] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *ICLR*, 2023. 5

[8] Laurent Condat and Saleh Mosaddegh. Joint demosaicking and denoising by total variation minimization. In *ICIP*, pages 2781–2784, 2012. 3

[9] Tobi Delbruck et al. Frame-free dynamic digital vision. In *Proc. of Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*, pages 21–26. Citeseer, 2008. 2, 3

[10] Yanchen Dong, Jing Zhao, Ruiqin Xiong, and Tiejun Huang. 3D residual interpolation for spike camera demosaicing. In *ICIP*, pages 1461–1465. IEEE, 2022. 2, 7, 8

[11] Yanchen Dong, Ruiqin Xiong, Jing Zhao, Jian Zhang, Xiaopeng Fan, Shuyuan Zhu, and Tiejun Huang. Joint demosaicing and denoising for spike camera. In *AAAI*, pages 1582–1590, 2024. 1, 2, 3, 7, 8

[12] Yanchen Dong, Ruiqin Xiong, Jing Zhao, Jian Zhang, Xiaopeng Fan, Shuyuan Zhu, and Tiejun Huang. Learning a deep demosaicing network for spike camera with color filter array. *IEEE TIP*, 2024. 1, 2, 6, 7, 8

[13] DC Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982. 8

[14] Omar A Elgendy, Abhiram Gnanasambandam, Stanley H Chan, and Jiaju Ma. Low-light demosaicking and denoising for small pixels using learned frequency selection. *IEEE TCI*, 7:137–150, 2021. 3

[15] Berthy T. Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L. Bouman, and William T. Freeman. Score-Based Diffusion Models as Principled Priors for Inverse Imaging. In *ICCV*, 2023. 2

[16] Eric R Fossum, Jiaju Ma, and Saleh Masoodian. Quanta image sensor: Concepts and progress. *Advanced Photon Counting Techniques X*, 9858:985805, 2016. 3

[17] Eric R Fossum, Jiaju Ma, Saleh Masoodian, Leo Anzagira, and Rachel Zizza. The quanta image sensor: Every photon counts. *Sensors*, 16(8):1260, 2016. 3

[18] Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand. Deep joint demosaicking and denoising. *ACM TOG*, 35(6):1–12, 2016. 3

[19] Felix Heide, Mushfiqur Rouf, Matthias B Hullin, Bjorn Labitzke, Wolfgang Heidrich, and Andreas Kolb. High-quality computational imaging through simple lenses. *ACM TOG*, 32 (5):1–14, 2013. 3

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 5

[21] 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, 2022. 2, 3

[22] Siddhant Jain, Daniel Watson, Eric Tabellion, Aleksander Hołyński, Ben Poole, and Janne Kontkanen. Video Interpolation with Diffusion Models. In *CVPR*, 2024. 2

[23] Daisuke Kiku, Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi. Beyond color difference: Residual interpolation for color image demosaicking. *IEEE TIP*, 25(3): 1288–1300, 2016. 3

[24] Filippos Kokkinos and Stamatios Lefkimmiatis. Deep image demosaicking using a cascade of convolutional residual denoising networks. In *ECCV*, pages 303–319, 2018. 3

[25] Juan Antonio Leñero-Bardallo, Teresa Serrano-Gotarredona, and Bernabé Linares-Barranco. A 3.6 μs latency asynchronous frame-free event-driven dynamic-vision-sensor. *IEEE Journal of Solid-State Circuits*, 46(6):1443–1455, 2011. 2, 3

[26] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. VidToMe: Video token merging for zero-shot video editing. In *CVPR*, pages 7486–7495, 2024. 8

[27] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 dB 15μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. 2, 3

[28] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *ECCV*, pages 430–448. Springer, 2024. 6

[29] Lin Liu, Xu Jia, Jianzhuang Liu, and Qi Tian. Joint demosaicing and denoising with self guidance. In *CVPR*, pages 2240–2249, 2020. 3

[30] Yunfan Lu, Yijie Xu, Wenzong Ma, Weiyu Guo, and Hui Xiong. Event camera demosaicing via swin transformer and pixel-focus loss. In *CVPR*, pages 1095–1105, 2024. 3

[31] Sizhuo Ma, Varun Sundar, Paul Mos, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Seeing photons in color. *ACM TOG*, 42(4):1–16, 2023. 2, 3

[32] Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *IEEE TIP*, 17(1): 53–69, 2007. 3

[33] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *ICCV*, pages 2272–2279, 2009.

[34] Henrique S Malvar, Li-wei He, and Ross Cutler. High-quality linear interpolation for demosaicing of bayer-patterned color images. In *ICASSP*, pages iii–485, 2004. 3

[35] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 8

[36] Ibrahim Pekkucuksen and Yucel Altunbasak. Gradient based threshold free color filter array interpolation. In *ICIP*, pages 137–140, 2010. 3

[37] Reza Pournaghi and Xiaolin Wu. Coded acquisition of high frame rate video. *IEEE TIP*, 23(12):5670–5682, 2014. 2

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 6

[39] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. XVFI: extreme video frame interpolation. In *ICCV*, 2021. 6

[40] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3

[41] Daniel Stanley Tan, Wei-Yang Chen, and Kai-Lung Hua. DeepDemosaicking: Adaptive image demosaicking via multiple deep fully convolutional networks. *IEEE TIP*, 27(5): 2408–2419, 2018. 3

[42] Runjie Tan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Color image demosaicking via deep residual learning. In *ICME*, pages 793–798, 2017. 3

[43] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. In *ICLR*, 2022. 2

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 8

[45] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Marc Levoy, and Mark Horowitz. High-speed videography using a dense camera array. *CVPR*, 2004. 2

[46] Wenzhu Xing and Karen Egiazarian. End-to-end learning for joint image demosaicing, denoising and super-resolution. In *CVPR*, pages 3507–3516, 2021. 3

[47] Senyan Xu, Zhijing Sun, Jiaying Zhu, Yurui Zhu, Xueyang Fu, and Zheng-Jun Zha. Demosaicformer: Coarse-to-fine demosaicing network for hybridevs camera. In *CVPR*, pages 1126–1135, 2024. 3

[48] Siqi Yang, Zhaojun Huang, Yakun Chang, Bin Fan, Zhaofei Yu, and Boxin Shi. Real-data-driven 2000 FPS color video from mosaicked chromatic spikes. In *ECCV*, 2024. 2, 3, 6, 7, 8

[49] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling Up to Excellence: Practicing Model Scaling for Photo-Realistic Image Restoration In the Wild. In *CVPR*, 2024. 2

[50] Guoshen Yu, Guillermo Sapiro, and Stéphane Mallat. Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity. *IEEE TIP*, 21 (5):2481–2499, 2011. 3

[51] Jie Zhang, Jonathan Newman, Zeguan Wang, Yong Qian, Pedro Feliciano-Ramos, Wei Guo, Takato Honda, Zhe Sage Chen, Changyang Linghu, Ralph Etienne-Cummings, et al. Pixel-wise programmability enables dynamic high-snr cameras for high-speed microscopy. *Nature Communications*, 15 (1):4480, 2024. 2

[52] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE TIP*, 24(8): 2579–2591, 2015. 8

[53] Zhao Zhang, Huan Zheng, Richang Hong, Mingliang Xu, Shuicheng Yan, and Meng Wang. Deep color consistent network for low-light image enhancement. In *CVPR*, pages 1899–1908, 2022. 3

[54] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2Imgnet: Learning to reconstruct dynamic scene from continuous spike stream. In *CVPR*, pages 11996–12005, 2021. 2

[55] Jing Zhao, Ruiqin Xiong, Jiyu Xie, Boxin Shi, Zhaofei Yu, Wen Gao, and Tiejun Huang. Reconstructing clear image for high-speed motion scene with a retina-inspired spike camera. *IEEE TCI*, 8:12–27, 2021. 2

[56] Yajing Zheng, Lingxiao Zheng, Zhaofei Yu, Boxin Shi, Yonghong Tian, and Tiejun Huang. High-speed image reconstruction through short-term plasticity for spiking cameras. In *CVPR*, pages 6358–6367, 2021. 2

[57] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *ICME*, pages 1432–1437, 2019. 2, 4, 7, 8

[58] Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. Retina-like visual image reconstruction via spiking neural model. In *CVPR*, pages 1438–1446, 2020. 2

SpikeDiff: Zero-shot High-Quality Video Reconstruction from Chromatic Spike Camera and Sub-millisecond Spike Streams

Supplementary Material

Siqi Yang[1,2,3] Jinxiu Liang[2,3,4*] Zhaojun Huang[2,3] Yeliduosi Xiaokaiti[2,3]
Yakun Chang[5,6] Zhaofei Yu[1,3] Boxin Shi[2,3,1*]

[1] Institute for Artificial Intelligence, Peking University

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

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

[4] National Institute of Informatics [5] Institute of Information Science, Beijing Jiaotong University

[6] Visual Intelligence +X International Cooperation Joint Laboratory of the Ministry of Education

cssherryliang@gmail.com, shiboxin@pku.edu.cn

6. Working mechanism of spike camera

As introduced in Sec. 3.1, the spike camera captures the scene in a continuous accumulation and trigger mechanism. We also describe this working mechanism with a finite state automaton (FSA), as shown in Figure 6. Each pixel in the spike camera asynchronously accumulates the incoming photons and readout the triggered spikes at a high sampling rate (*e.g.*, $20,000$ Hz). Spike pixel starts with an initial voltage $E = 0$, and accumulates the incoming photons ΔI during the last period (*e.g.*, $1/20000$ s) with a conversion ratio α. If the accumulated voltage $E + \alpha\Delta I$ exceeds the pre-defined threshold E_{th}, the pixel will trigger a spike (readout 1) and reset the voltage. Otherwise, the pixel will not trigger a spike (readout 0) and keep the accumulated voltage.

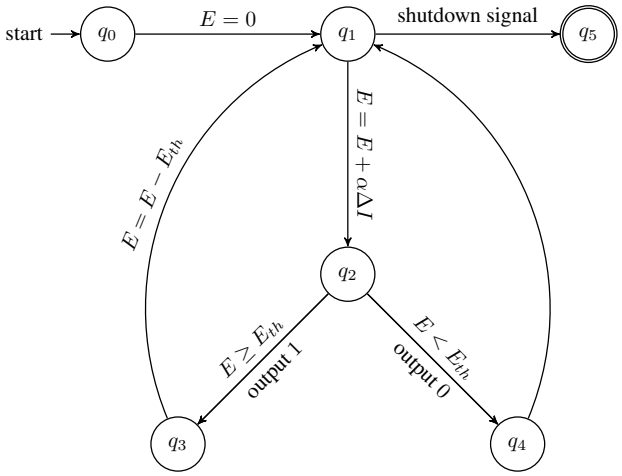


Figure 6. Spike camera working mechanism described using finite state automaton, where E denotes the accumulated voltage, E_{th} denotes the voltage threshold, α denotes the conversion ratio, and ΔI denotes the incoming photons during last accumulation period.

7. Additional implementation details

The results of the compared methods, except for CSpkNet [3], are produced using the codes and checkpoints provided by their authors. For CSpkNet [3], since only the code is obtained, we retrained following the original paper with synthetic dataset. We used the same spike model as these methods in simulation to ensure fairness. Note that TFI and TFP are not learning-based methods. Instead of starting from the completely random initialization $\boldsymbol{Z}_T$, we utilize an intermediate latent state to accelerate the diffusion process, which is widely used in diffusion pipelines:

$$\boldsymbol{Z}_{t_s} = \sqrt{\bar{\alpha}_{t_s}}\mathcal{E}(\boldsymbol{Y}) + \sqrt{1-\bar{\alpha}_{t_s}}z, \quad z \sim \mathcal{N}(0, I). \quad (19)$$

The hyperparameter k controls the smoothness of soft quantization (higher k approaches harder quantization). We empirically select $k = 50$ in experiments for best reconstruction quality, as shown in Tab. 4.

To collect the real-captured chromatic spike dataset for qualitative evaluation, we use Spike M1K40-H2-Gen3 (chromatic version) from SpikeSee[1], which captured Bayer-pattern spike streams at $20,000$ Hz, with a spatial resolution of 1000×1000, as shown in Figure 7.

8. Additional experiments results

8.1. Additional qualitative results

Real-captured chromatic spikes. We conduct additional experiments on real-world captured chromatic spikes to analyze the performance of our proposed method qualitatively. As shown in Figure 8(a), we spin the umbrella with rainbow colors in front of the light source. Our proposed SpikeDiff

* Corresponding authors.

[1] https://www.spikesee.com/product.html

Figure 7. Spike M1K40-H2-Gen3 (chromatic version) camera. We use this camera to capture spike streams for evaluation on real data.

recovers the accurate rainbow colors with clean and sharp textures, leading to apparently better visual quality than other methods. As for Figure 8(b), we capture the rotating fan before the color checkerboard. As the time interval of chromatic spikes is limited to 0.5 ms, all methods are free of motion blur. But most method suffer from noise perturbation or over-smoothing. In contrast, SpikeDiff recovers clean and high-quality frames. In Figure 8(c-d), we focus on another rotating fan with color tapes. SpikeDiff successfully recover the color, suppress the noise, and preserve the sharp edges in highlighted areas. In conclusion, the additional qualitative evaluation demonstrates the superiority of our proposed method over existing chromatic spike reconstruction methods, especially in terms of chromatic spikes from sub-millisecond time intervals.

Simulated chromatic spikes. As described in Sec. 4, we generate a synthetic chromatic spikes dataset from high-frame-rate videos to evaluate the performance of our proposed method quantitatively. We further visualize the reconstruction results of SpikeDiff and existing methods on the synthetic dataset in Figure 9, together with the ground truth frames. The results show that our proposed method can recover video frames with the best visual quality, demonstrating the effectiveness of our proposed method.

8.2. Analysis of the degradation operators

Visualization of degradation process. We provide detailed visualization of the degradation process in the calculation of chromatic spikes' likelihood. As shown in Figure 13, our proposed degradation operators, including mosaicking $\boldsymbol{M}$, color casting $\boldsymbol{C}$, and quantization $\mathcal{Q}$, gradually transform the sampled video frame $\boldsymbol{X}_{0|t}$ to the same distribution as SFR frames $\boldsymbol{Y}$. Firstly, the mosaicking operator degrades the colored image to a Bayer-patterned mosaic image, whose debayering result is illustrated for better visualization, producing similar color bleeding as the SFR estimations in the blur bounding boxes. Consequently, the color casting operator transforms the mosaic image to the color distribution of SFR frames. Finally, a soft quantization operator is applied to each pixel and conducts a quantization pattern similar to the SFR frames, as indicated by the red bounding boxes.

Table 4. Analysis of the hyperparameter k in soft quantization.

k	PSNR↑	SSIM↑	FID↓	NIQE↓	IL-NIQE↓
10	17.088	0.589	5.115	6.787	45.029
50	18.694	0.750	2.880	5.173	38.535
200	17.889	0.542	4.127	6.795	49.158

Table 5. Quantitative evaluation of our proposed method SpikeDiff, with and without multiscale enhancement.

Method	PSNR↑	SSIM↑	FID↓	NIQE↓	IL-NIQE↓
w/ Multiscale	18.694	0.750	2.880	5.173	38.535
w/o Multiscale	17.549	0.570	3.706	6.809	48.350

Handling of color casting. To further demonstrate the superiority of our proposed method over existing chromatic spike reconstruction methods, which do not consider color casting in their models, we adapt these methods to convert their final outputs to the desired color distribution with gray world assumption (the same as ours). As shown in Figure 10, introducing color casting to existing methods can slightly improve their visual quality, but the noise and artifacts in the results cannot be eliminated. And our simulated dataset is free of color casting effects, thereby eliminate the potential influence of color casting in quantitative evaluations.

8.3. Analysis of time intervals

SpikeDiff is the first zero-shot method that can recover high-quality video frames from noisy real-captured chromatic spikes, even with extremely limited time intervals, *e.g.*, sub-millisecond. All the existing deep learning-based methods require much more spikes (*e.g.*, $\geq$ 2ms) to leverage richer information and suppress the noise with motion estimation. However, most of these methods suffer from the inaccurate estimation of optical flow and imperfect noise modeling, producing unsatisfactory results even with longer time intervals. As shown in Tab. 6, our proposed SpikeDiff also outperforms existing methods with longer time intervals as their declarations among most of the metrics.

8.4. Analysis of multiscale enhancement

We conduct quantitative analysis on the effectiveness of multiscale enhancement in our proposed method. As shown in Table 5, the multiscale enhancement improves the performance of SpikeDiff, with only negligible 0.5G FLOPs increase.

8.5. Comparison to diffusion-based methods

We compare our proposed method with other diffusion-based methods [4, 7], by applying the pretrained image / video

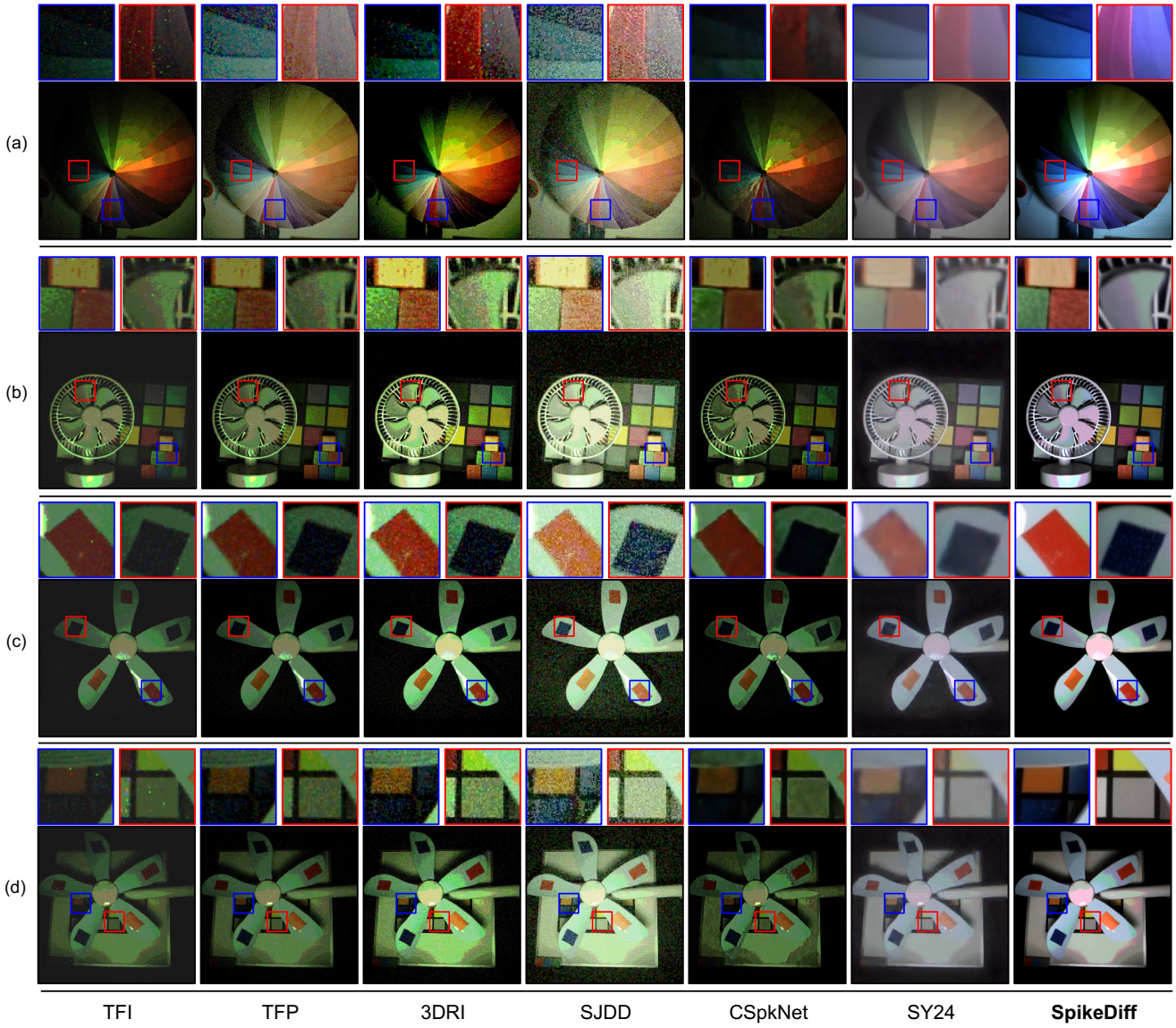


Figure 8. Additional qualitative comparison between our proposed SpikeDiff and existing reconstruction methods on real-captured chromatic spikes. All results are recovered from sub-millisecond chromatic spikes (0.5ms). Details in red / blue bounding boxes are shown on the top.

Table 6. Quantitative comparison of the proposed SpikeDiff (with 0.5ms time intervals) with existing chromatic spike reconstruction methods (with $\geq$ 2.0ms time intervals, satisfying the original declaration of each method). The best and second-best results are highlighted in **red** and blue, respectively.

Method	PSNR↑	SSIM↑	FID↓	NIQE↓	IL-NIQE↓
SpikeDiff (0.5ms)	18.694	**0.750**	**2.880**	**5.173**	**38.535**
SY24 [6] (3.0ms)	14.163	0.629	20.239	10.199	80.243
SJDD [2] (2.0ms)	11.250	0.505	7.843	7.855	41.079
3DRI [1] (2.0ms)	**21.618**	0.625	5.991	8.345	41.816
CSpkNet [3] (2.0ms)	14.393	0.744	3.473	**5.982**	40.066
TFP [8] (2.0ms)	13.429	0.449	14.986	13.089	60.693
TFI [8] (2.0ms)	16.924	0.700	3.449	11.556	49.377

restoration diffusion pipelines to the SFR frames $\boldsymbol{Y}$. As shown in Figure 12, the naive application of these diffusion models leads to unsatisfactory results, where the reconstructed images suffer from severe artifacts, *e.g.*, producing generated textures or suffering from quantization effects. In contrast, our proposed method effectively suppresses the generative artifacts, recovers the color information, and achieves a more visually pleasing result. This comparison demonstrates the effectiveness of video diffusion-based posterior estimation, which combines the existing diffusion pipeline with the external physics-based guidance from the spikes. Note that we integrate color casting as pre-processing for these methods to eliminate its potential influence. Compared to these methods, SpikeDiff leverages additional physics-based guidance from chromatic spikes via differentiable operators, avoiding the instability of applying techniques like token merging to spikes, particularly regarding optical flow dependencies. Instead, SpikeDiff achieves temporal

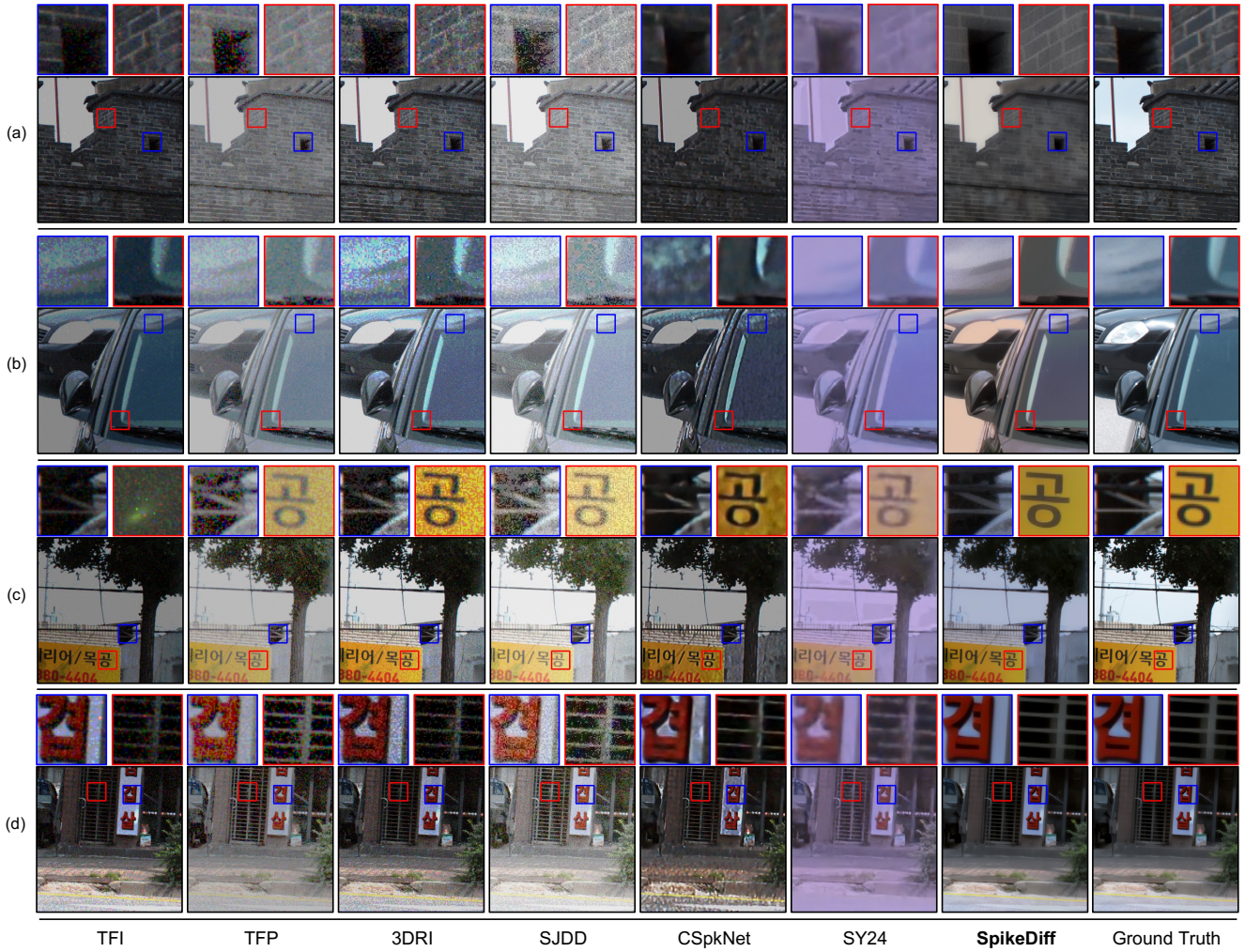


Figure 9. Visualization of the chromatic spike reconstruction results on synthetic dataset. The chromatic spikes are simulated from real-world high-frame-rate videos. All the results are reconstructed from sub-millisecond chromatic spikes (0.5ms). SpikeDiff achieves the best visual quality and texture preserving among all the methods. Details in red / blue bounding boxes are shown on the top.

consistency by leveraging high-fidelity reconstruction from time-continuous chromatic spikes.

9. Further discussion

9.1. Alternative SFR estimation

As we introduced in Sec. 3, our proposed method SpikeDiff starts from the spike firing rate (SFR) estimations $\boldsymbol{Y}$, which is perturbed by spike noise, integrate SFR frames into the diffusion-based posterior sampling process via chromatic spikes' likelihood estimation, and finally recover high-quality video frames from these SFR frames. Despite the SFR estimation method we used in Eq. 4 (TFP [8]), there is another method (TFI [8]) which calculates the firing interval between two adjacent spikes and then takes its reciprocal as the estimation of firing rate:

$$\boldsymbol{Y}'(i,\tau) = 1/(\tau_{\text{next}} - \tau_{\text{last}}), \tag{20}$$

$$\tau_{\text{next}} = \min\{\tau' > \tau | S(i,\tau') = 1\}, \tag{21}$$

$$\tau_{\text{last}} = \max\{\tau' < \tau | S(i,\tau') = 1\}. \tag{22}$$

However, as shown in Fig. 4, the noise contamination of TFI does not follow the same pattern as TFP, which cannot be assumed as a Gaussian distribution. In experiments, we demonstrate that directly replacing $\boldsymbol{Y}$ with $\boldsymbol{Y}'$ in SpikeDiff leads to significant artifacts in generated video frames, highly related to the noisy pixels in the SFR frames, which is consistent with our hypothesis, as shown in Figure 14. Therefore, our proposed method is not compatible with TFI estimations. We believe it requires additional noise modeling and optimization designs to integrate TFI into diffusion-based posterior sampling, due to its out-of-distribution noise characteristics.

Figure 10. Qualitatively comparison between our proposed SpikeDiff and other chromatic spike reconstruction methods, with color casting based on gray world assumption as post-processing for other methods. Compared to Fig. 4 and Figure 8, the correction of color distribution slightly improves the visual quality of other reconstruction methods, but cannot eliminate any noise or artifacts.

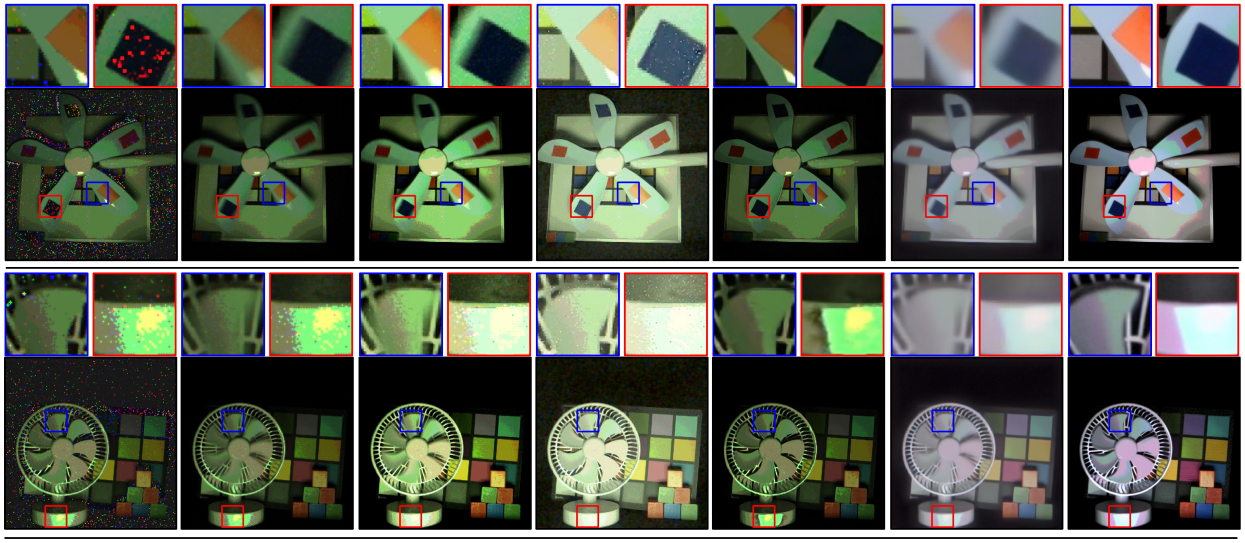


Figure 11. Qualitatively comparison between our proposed SpikeDiff (with 0.5ms time intervals) and other chromatic spike reconstruction methods (with $\geq$ 2.0ms time intervals, satisfying the original declaration of each method). With longer time intervals, existing methods either suffer from motion blur or residuary noisy pixels. Our proposed SpikeDiff recovers the most clean and visually pleasant reconstruction results even with sub-millisecond spikes.

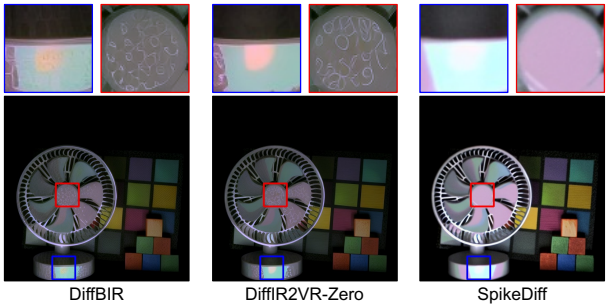


Figure 12. Comparison to image/video restoration diffusion models, *i.e.*, DiffBIR [4], DiffIR2VR-Zero [7]. Although color casting can be integrated as pre-processing, directly application of these diffusion-based image/video restoration methods still suffers from quantization and serious generation artifacts, while SpikeDiff can produce high-quality results faithful to the chromatic spikes.

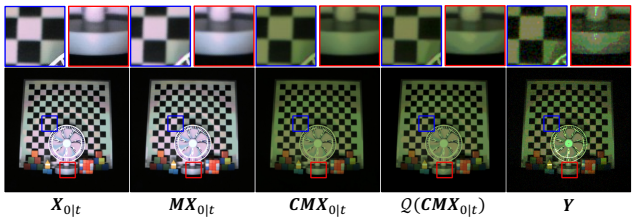


Figure 13. Detailed visualization of the degradation process. The blue bounding boxes show the effects of our mosaicking operator, and the red bounding boxes demonstrates the effectiveness of our soft quantization operator.

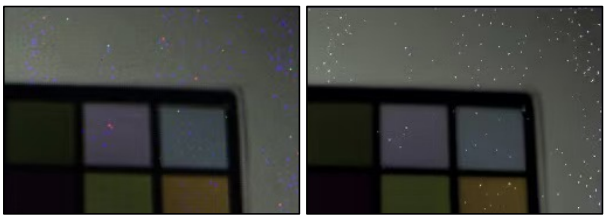


Figure 14. TFI-based SFR estimation and the corresponding result from adapted SpikeDiff pipeline. The white noisy points in recovered frames are caused by the out-of-distribution noise from TFI-based SFR estimation.

9.2. Further reduced time-intervals

Our proposed method SpikeDiff already achieves high-quality reconstruction from sub-millisecond chromatic spikes, and outperforms existing methods with much longer spike streams as input (*e.g.*, 2.0 to 3.0 ms), as demonstrated in Figure 8 and Table 6. Beyond of this, we also conduct experiments on further reduced time intervals, *e.g.*, 0.1 ms, equivalent to only 2 spike frames. However, due to the missing of texture information and perturbation of noisy spikes in extremely limited time intervals, even our proposed method still cannot recover clean frames from such input.

9.3. Inference speed analysis

Integrating pretrained diffusion models into chromatic spike reconstruction problem provides principled priors to eliminate the potential spike noise, but it also requires a large amount of computation resources. In our experiments, we

Table 7. Offline inference speed of SpikeDiff and other methods to recover a video frame from chromatic spikes, benchmarked with Intel i9-12900K and NVIDIA RTX3090, averaged over 50 runs.

Method	YS24	3DRI	SJDD	CSpkNet	SpikeDiff	Baseline
Runtime (s)	0.07	1.25	3.22	0.51	20	12.5

Table 8. FLOPs of SpikeDiff and other methods to recover a single frame from chromatic spikes.

Method	YS24	3DRI	SJDD	CSpkNet	SpikeDiff
TFLOPs	0.65	12.38	30.30	4.53	182.72

compare the inference speed and floating point operations of SpikeDiff with other chromatic spike reconstruction methods, as shown in Table 7 and Table 8. The inference speed of our proposed method is slower, but we believe it is acceptable for offline processing tasks, and SpikeDiff achieves zero-shot reconstruction with a dominant performance in terms of extremely short time interval. Additionally, SpikeDiff can leverage accelerating techniques from diffusion models, *e.g*. DeepCache [5], which can mitigate this problem but is beyond our scope. And diffusion techniques such as mask-shift sampling can also be integrated to SpikeDiff, which can improve the spatial resolution.

References

[1] Yanchen Dong, Jing Zhao, Ruiqin Xiong, and Tiejun Huang. 3D residual interpolation for spike camera demosaicing. In *ICIP*, pages 1461–1465. IEEE, 2022. 3

[2] Yanchen Dong, Ruiqin Xiong, Jing Zhao, Jian Zhang, Xiaopeng Fan, Shuyuan Zhu, and Tiejun Huang. Joint demosaicing and denoising for spike camera. In *AAAI*, pages 1582–1590, 2024. 3

[3] Yanchen Dong, Ruiqin Xiong, Jing Zhao, Jian Zhang, Xiaopeng Fan, Shuyuan Zhu, and Tiejun Huang. Learning a deep demosaicing network for spike camera with color filter array. *IEEE TIP*, 2024. 1, 3

[4] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *ECCV*, pages 430–448. Springer, 2024. 2, 6

[5] Xinyin Ma, Gongfan Fang, and Xinchao Wang. DeepCache: Accelerating diffusion models for free. In *CVPR*, pages 15762–15772, 2024. 7

[6] Siqi Yang, Zhaojun Huang, Yakun Chang, Bin Fan, Zhaofei Yu, and Boxin Shi. Real-data-driven 2000 FPS color video from mosaicked chromatic spikes. In *ECCV*, 2024. 3

[7] Chang-Han Yeh, Chin-Yang Lin, Zhixiang Wang, Chi-Wei Hsiao, Ting-Hsuan Chen, Hau-Shiang Shiu, and Yu-Lun Liu. DiffIR2VR-Zero: Zero-shot video restoration with diffusion-based image restoration models. *arXiv preprint arXiv:2407.01519*, 2024. 2, 6

[8] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *ICME*, pages 1432–1437, 2019. 3, 4