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.

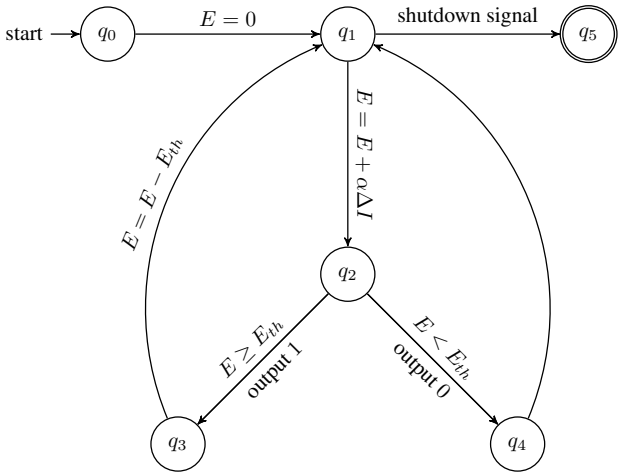

Figure 6. Spike camera working mechanism described using finite state automaton, where E denotes the accumulated voltage, E_{th} denotes the voltage threshold, α denotes the conversion ratio, and ΔI denotes the incoming photons during last accumulation period.

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

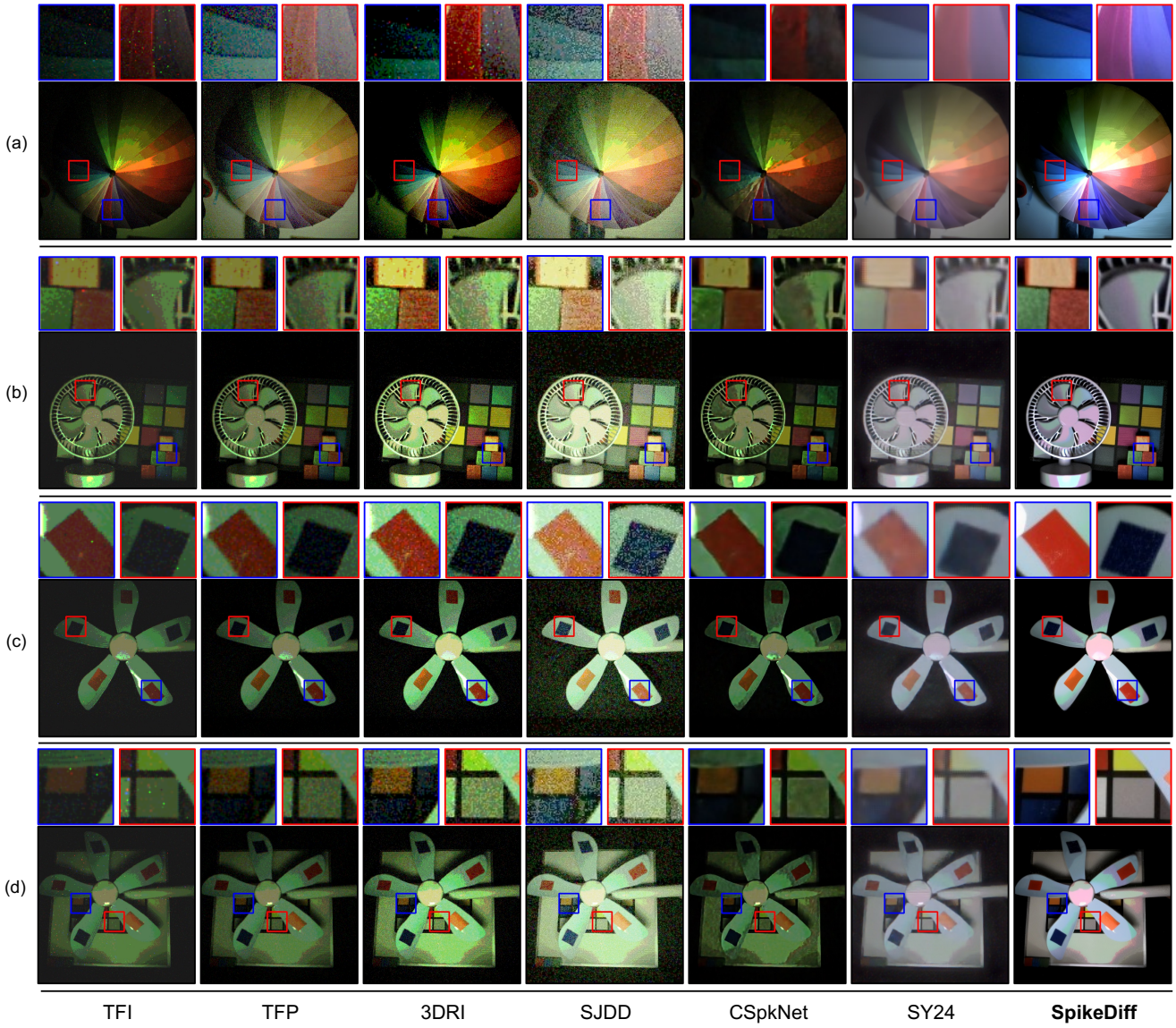

Figure 8. Additional qualitative comparison between our proposed SpikeDiff and existing reconstruction methods on real-captured chromatic spikes. All results are recovered from sub-millisecond chromatic spikes (0.5ms). Details in red / blue bounding boxes are shown on the top.

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

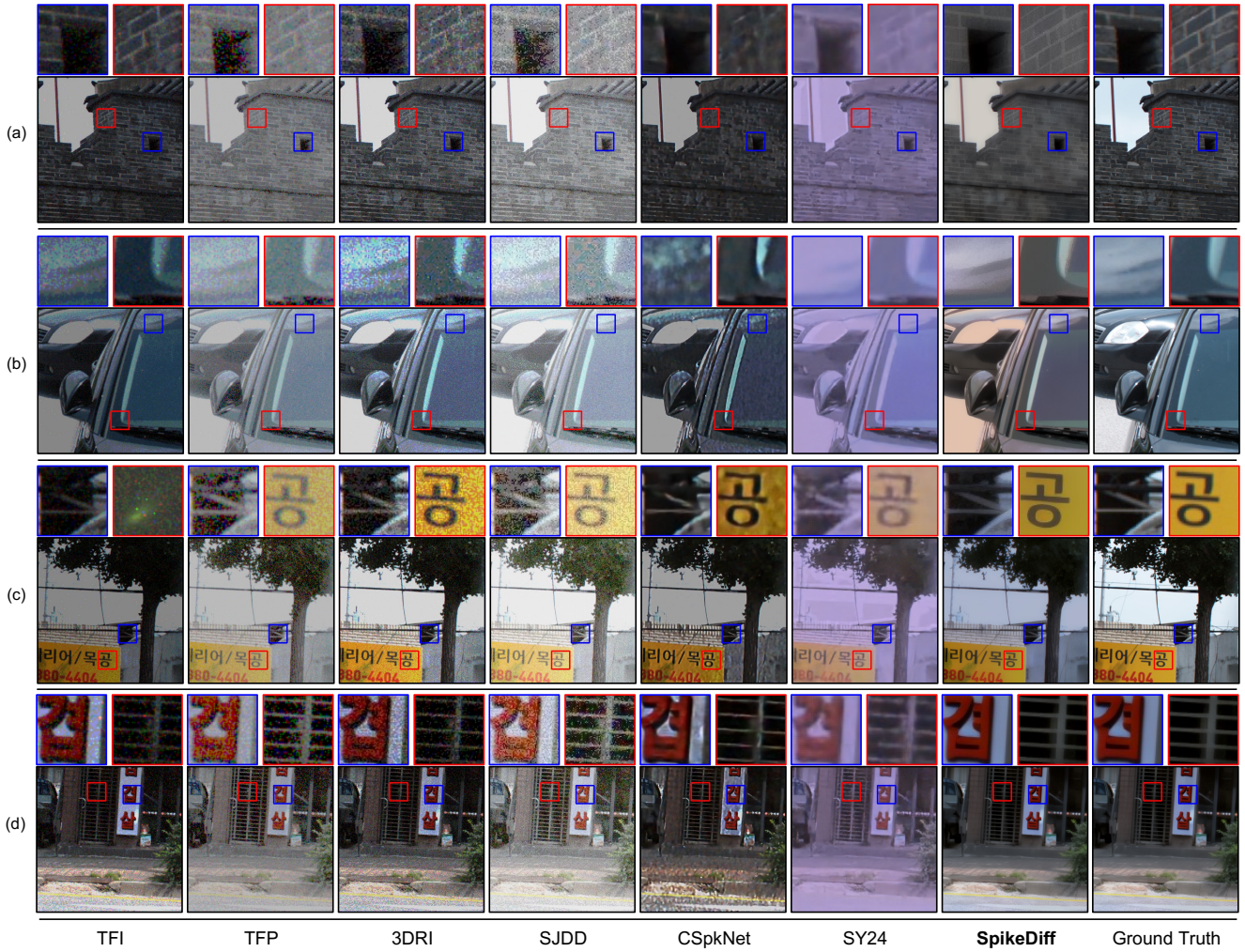

Figure 9. Visualization of the chromatic spike reconstruction results on synthetic dataset. The chromatic spikes are simulated from real-world high-frame-rate videos. All the results are reconstructed from sub-millisecond chromatic spikes (0.5ms). SpikeDiff achieves the best visual quality and texture 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.

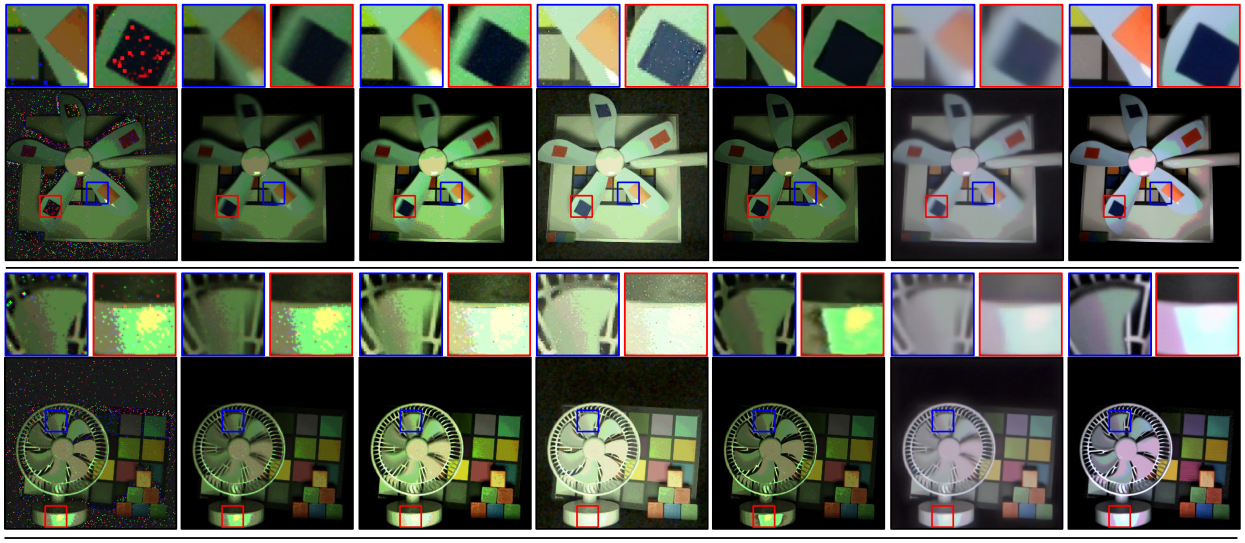

Figure 11. Qualitatively comparison between our proposed SpikeDiff (with 0.5ms time intervals) and other chromatic spike reconstruction methods (with $\geq$ 2.0ms time intervals, satisfying the original declaration of each method). With longer time intervals, existing methods either suffer from motion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

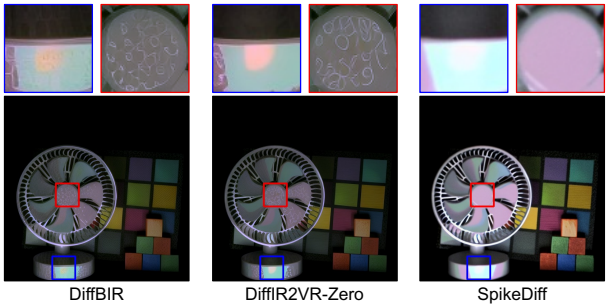

Figure 12. Comparison to image/video restoration diffusion models, *i.e.*, DiffBIR [4], DiffIR2VR-Zero [7]. Although color casting can be integrated as pre-processing, directly application of these diffusion-based image/video restoration methods still suffers from quantization and serious generation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

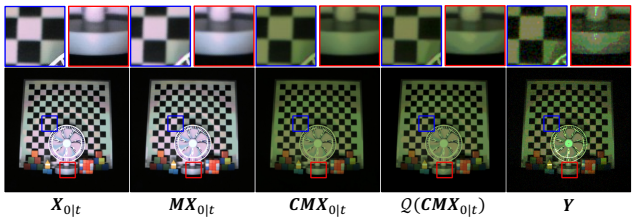

Figure 13. Detailed visualization of the degradation process. The blue bounding boxes show the effects of our mosaicking operator, and the red bounding boxes demonstrates the effectiveness of our soft quantization operator.

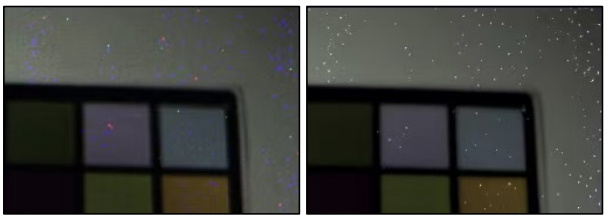

Figure 14. TFI-based SFR estimation and the corresponding result from adapted SpikeDiff pipeline. The white noisy points in recovered frames are caused by the out-of-distribution noise from TFI-based SFR estimation.

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