

박사학위논문
Ph.D. Dissertation

이벤트 카메라 기반의 강건한 영상 복원 기술

Event-guided Robust Image and Video Restoration

2026

정준식 (鄭準植 Jung, Junsik)

한국과학기술원

Korea Advanced Institute of Science and Technology

박 사 학 위 논 문

이벤트 카메라 기반의 강건한 영상 복원 기술

2026

정 준 식

한 국 과 학 기 술 원

전산학부

이벤트 카메라 기반의 강건한 영상 복원 기술

정 준 식

위 논문은 한국과학기술원 박사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2026년 6월 10일

심사위원장 윤 성 의 (인)

심 사 위 원 김 민 혁 (인)

심 사 위 원 김 태 균 (인)

심 사 위 원 이 재 길 (인)

심 사 위 원 최 성 희 (인)

Event-guided Robust Image and Video Restoration

Junsik Jung

Advisor: Sung-Eui Yoon

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

Daejeon, Korea
June 19, 2026

Approved by

Sung-Eui Yoon
Professor of School of Computing

The study was conducted in accordance with Code of Research Ethics¹.

¹ Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

DCS

정준식. 이벤트 카메라 기반의 강건한 영상 복원 기술. 전산학부. 2026년.
53+iii 쪽. 지도교수: 윤성의. (영문 논문)

Junsik Jung. Event-guided Robust Image and Video Restoration. School of
Computing. 2026. 53+iii pages. Advisor: Sung-Eui Yoon. (Text in English)

초 록

기존 프레임 기반 카메라는 동적 환경에서 모션 블러 등의 화질 열화에 근본적으로 취약하다. 기존 복원 기법들이 벤치마크에서는 우수한 성능을 보이나, 극심한 도메인 차이, 독립적인 이미지 모델의 시간적 불일치, 예측 불가능한 실제 촬영 조건을 반영하지 못하는 설계적 한계로 인해 실제 환경으로의 일반화에는 어려움을 겪는다. 본 학위논문은 이러한 한계를 극복하고자 이벤트 카메라의 높은 시간 해상도를 활용한 강건한 영상 복원 기법들을 제안한다. 첫째, 복잡한 시간적 제약 조건을 암시적으로 학습하여, 동적 노출 환경에서 비디오 프레임 보간과 모션 디블러링을 동시에 수행하는 이벤트 기반 프레임 복원 프레임워크를 제안한다. 둘째, 디블러링의 도메인 편향 문제를 해결하기 위한 이벤트 기반 블러 합성 프레임워크를 제안한다. 이벤트 데이터로 영상의 콘텐츠와 모션을 분리하여 타겟 도메인에 최적화된 학습 데이터를 합성함으로써, 기존 모델들의 강건한 미세 조정을 지원한다. 셋째, 사전 학습된 이미지 확산 모델을 비디오에 적용할 때 발생하는 시간적 불안정성을 완화하는 제로샷 비디오 디블러링 프레임워크를 제안한다. 공간 적응형 시간 가이드를 도입해 동적 및 정적 영역별 맞춤형 제약을 가함으로써, 추가 학습 없이도 이미지 확산 모델의 강력한 사전 지식을 보존하며 프레임 간 일관성을 확보한다. 광범위한 실험을 통해 제안 방법론들의 우수성을 입증하였으며, 본 연구는 미지의 실제 환경에서도 강건하고 도메인 적응적이며 시간적 일관성을 보장하는 이벤트 기반 영상 복원 기술의 발전에 기여한다.

핵심 낱말 계산 사진술, 이벤트 카메라, 프레임 보간, 움직임 블러 제거

Abstract

Conventional frame-based cameras fundamentally suffer from severe degradations, such as motion blur, when capturing dynamic environments. While existing restoration methods have achieved remarkable success on standard benchmarks, they often struggle to generalize to real-world challenges. These limitations primarily stem from severe domain shifts, the inherent temporal inconsistency of independent image models, and restrictive model designs that fail to account for complex, unknown capturing conditions. To address these challenges, this dissertation proposes a robust event-guided image and video restoration framework that leverages the high temporal resolution of event cameras. First, we introduce an event-guided framework for joint video frame interpolation and motion deblurring under dynamic exposures, achieving superior performance by implicitly modeling complex temporal constraints. Second, to tackle domain shifts in motion deblurring, we propose an event-guided blur synthesis framework. By decoupling content from motion, it synthesizes diverse target-domain training pairs for robust model fine-tuning. Lastly, we propose a zero-shot video deblurring framework that mitigates temporal inconsistency in pre-trained Image Diffusion Models (IDMs). Using event-guided spatially-adaptive temporal guidance, we enforce consistency across both dynamic and static regions, exploiting strong IDM priors without fine-tuning. Extensive experiments demonstrate the effectiveness of our methods, contributing to the development of event-guided restoration frameworks that are robust under real-world scenarios.

Keywords Computational Photography, Event Camera, Video Frame Interpolation, Motion Deblurring

Contents

Contents	i
List of Tables	ii
List of Figures	iii
Chapter 1. Introduction	1
Chapter 2. Learning Event-guided Exposure-agnostic Video Frame Interpolation via Adaptive Feature Blending	3
2.1 Introduction	3
2.2 Related Work	4
2.3 Method	5
2.4 Experiments	8
2.5 Conclusion	10
2.6 Appendix	11
Chapter 3. EvBS: Event-guided Blur Synthesis for Domain-adaptive Motion Deblurring	19
3.1 Introduction	19
3.2 Related Work	20
3.3 Method	21
3.4 Experiments	27
3.5 Conclusion	32
3.6 Appendix	32
Chapter 4. EvZVD: Event-guided Zero-shot Video Deblurring Using Pre-trained Image Deblurring Diffusion Model	37
4.1 Introduction	37
4.2 Related Work	38
4.3 Method	38
4.4 Experiments	42
4.5 Conclusion	44
Chapter 5. Conclusion	45
Bibliography	46

List of Tables

2.1	Summary of feature comparisons among different video deblurring methods. Our method simultaneously satisfies all the key properties.	8
2.2	Quantitative results on GoPro and HighREV (10↓). Bold and <u>underlined</u> indicate the best and second-best scores, respectively.	9
2.3	Quantitative results on GoPro and HighREV (16↓). Bold and <u>underlined</u> indicate the best and second-best scores, respectively.	10
2.4	Ablation study on GoPro-10↓ (RandEx). All ablations are retrained from scratch, except Swap ω_τ	13
2.5	tOF↓ / tLP↓ evaluations on GoPro and HighREV with RandEX setting. Bold indicates the best scores.	13
2.6	NIQE evaluations on RealBlur-DAVIS. Bold indicates the best score.	14
2.7	Computational cost on GoPro-10↓ (RandEx setting). Runtime measured on a TITAN RTX GPU. Bold and <u>underlined</u> indicate the best and second-best results, respectively.	14
3.1	Quantitative results across deblurring benchmarks. We compare baseline performance with results obtained after adaptation via our method (w/ Ours). * indicates event-based models.	28
3.2	Comparison with recent domain-adaptive deblurring methods. Best results in bold	28
3.3	Effectiveness of motion components. Combining both intrinsic and extrinsic motion yields the highest PSNR. Conversely, applying random motion offsets severely degrades performance, validating the necessity of our motion transfer algorithm.	31
3.4	Computational cost analysis (adapting on EVRB with MAENet, using RTX 2080 Ti).	34
4.1	Quantitative evaluation on REVD and EVRB datasets. The best and second-best results are highlighted in bold and <u>underlined</u> , respectively. Our method significantly improves the temporal consistency (tOF, tLP, WE) of the pre-trained IDM (DeblurDiff) while preserving its high spatial restoration quality (PSNR, SSIM).	42
4.2	Ablation study on the effectiveness of different guidance components evaluated on the REVD dataset. CA: Cross-frame Attention, NCG: Noise Consistency Guidance, PCG: Prediction Consistency Guidance.	44

List of Figures

1.1	Comparison of visual data acquisition. A conventional RGB camera captures discrete frames with a fixed shutter period (left), whereas an event camera continuously records the spatiotemporal trajectory of moving objects (right).	1
2.1	Schematic comparison with a prior event-guided exposure-agnostic VFI method. (a) EBFI [1] restores sharp frames from a single blurry input via event modulation, but degrades as the target timestamp deviates due to the lack of temporal constraints. (b) Our method addresses this by explicitly leveraging temporal constraints via adaptive feature blending. (c) PSNR comparison on GoPro [2] shows that our method maintains more stable performance across varying timestamps.	4
2.2	Overview of our framework. Given blurry frames (I_0, I_1) with unknown exposures (T_{e_0}, T_{e_1}) , stacked events $\mathbb{E}_N$ over $2T$, and target timestamp τ , the model reconstructs the target frame $\hat{I}_\tau$. TES samples events around τ and unknown exposure, which are fused with frames. TIM generates an importance map ω_τ to adaptively blend the fused features (F_0, F_1) into F_τ , which is then decoded into $\hat{I}_\tau$. Shared encoders are color-coded.	5
2.3	Architectures of the TES (Sec. 2.3.2) and TIM (Sec. 2.3.3) modules.	6
2.4	Qualitative results on RealBlur-DAVIS (real blur & events)	11
2.5	Channel-wise importance score $\frac{\partial \mathcal{E}_{\tau \rightarrow [t_s, t_e]}}{\partial \mathbb{E}_N}$ indicating the sampling contribution of each time index. TES effectively selects events near the unknown exposure T_e (green) and target timestamp τ (blue), regardless of whether τ lies outside the exposure window.	12
2.6	Visualization of importance maps ω_τ for $\tau \in 3, 7, 12, 15, 25$, with symmetric exposure durations ($m = 7$) for I_0 and I_1	12
2.7	Channel-wise average importance scores $\frac{\partial \mathcal{E}_{\tau \rightarrow [t_s, t_e]}}{\partial \mathbb{E}_N}$ showing how much each temporal slice in $\mathbb{E}_N$ contributes to the sampled events. (a) Without the sampling process and (b) without the target timestamp τ . Higher scores indicate stronger selection by the TES module.	14
2.8	PSNR comparison over time with and without temporal constraints. Our method (red) maintains stable performance across target target timestamps, while the version without temporal constraints (blue & green) shows degradation.	14
2.9	Detailed architecture of the encoder and decoder. Frame and event encoders share the same structure. The model follows a U-Net design with three encoders and two decoders, where only the final decoder includes a refinement block.	15
2.10	Qualitative results on GoPro (synthetic blur & events).	15
2.11	Qualitative results on HighREV (synthetic blur & real events).	16
2.12	More qualitative results on RealBlur-DAVIS (real blur & events)	17
2.13	(continued) More qualitative results on RealBlur-DAVIS.	18

3.1	Illustration of EvBS. In the target domain representation, shapes denote visual content and colors represent distinct motion patterns. By leveraging events to explicitly decouple motion from content, EvBS enables both intrinsic blur synthesis (applying a content source’s inherent motion) and extrinsic blur synthesis (transferring motion from a separate motion source). This strategy generates a diverse set of synthesized training pairs within the target domain, facilitating enhanced model calibration.	20
3.2	Overview of the proposed Event-guided Blur Synthesis (EvBS) framework. Given test frames and their corresponding event streams from the target domain, the Dual Source Extractor (DSE) extracts pseudo-sharp patches as content sources and blurred patches as motion sources. The Blur Conditioning Module (BCM) then generates diverse blur conditions via two pathways: the Intrinsic-blur Condition Generator (InCG) utilizes the inherent motion of the content sources, while the Extrinsic-blur Condition Generator (ExCG) transfers motions from the motion sources to different content sources. These conditions guide a blurring model to synthesize a diverse set of blurred patches. Finally, these synthesized patches are used to fine-tune a pre-trained deblurring model, with the original content sources serving as direct supervision (red dashed line).	22
3.3	Illustration of the Dual Source Extractor. The blur score map $\mathcal{B}$ is computed from the event count map $\mathcal{C}_E$ and the blur magnitude map $\mathcal{M}$. Based on this map, sliding-window detection identifies content sources (low scores) and motion source candidates (high scores). Subsequent non-maximum suppression (NMS) and thresholding steps are utilized to eliminate overlapping patches and reduce redundancy. Motion sources are further filtered by the motion source selector based on motion consistency.	23
3.4	Intrinsic-blur Condition Generator. Given events $\mathbf{E}_C$ from the temporal window $[t-2, t+2]$ around a content source at frame t , we first estimate bidirectional optical flows. These flows are aggregated into a flow map and then multiplied by an event mask $\mathcal{M}$ to produce the final intrinsic flow map $\mathcal{F}_I$. Concurrently, $\mathbf{E}_C$ is processed via Flow-Enhanced Deviation Accumulation (FEDA) to generate the intrinsic event representation $\mathcal{E}_I$. Together, the pair $(\mathcal{F}_I, \mathcal{E}_I)$ serves as the intrinsic-blur conditions.	24
3.5	Extrinsic-blur Condition Generator. Flow maps $\mathcal{F}_C$ and $\mathcal{F}_M$ are first extracted from content events $\mathbf{E}_C$ and motion events $\mathbf{E}_M$. If their angular difference falls within $[\tau_{low}, \tau_{high}]$, the motion pattern from $\mathbf{E}_M$ is transferred to $\mathbf{E}_C$ via Algorithm 1 to produce the warped events $\tilde{\mathbf{E}}_C$. Finally, $\tilde{\mathbf{E}}_C$ is processed via FEDA to yield the extrinsic flow map $\mathcal{F}_E$ and event representation $\mathcal{E}_E$. The resulting condition $(\mathcal{F}_E, \mathcal{E}_E)$ effectively integrates the content’s spatial structure with external motion patterns.	25
3.6	Qualitative results across diverse models and datasets. We compare the blurry input (blur), pre-trained baseline (baseline), our adapted model (ours), and ground truth (sharp), with red boxes denoting zoomed-in regions. Our approach consistently restores finer details across both event-based deblurring models. Additional visual results for the other models (EBFI, NAFNet, MIMO-UNet) are shown in the supplementary material.	29
3.7	Visual comparison of deblurring results against recent domain-adaptive methods [3,4] Our method consistently yields sharper and more visually pleasing results across diverse scenes.	30

3.8	Qualitative examples of intrinsic and extrinsic blur synthesis. For each row, we display the content source and its flow map $\mathcal{F}_C$ (left), followed by the intrinsic synthesized blur . To demonstrate motion transfer, we show a separate motion source and its flow map $\mathcal{F}_M$ (middle), and the resulting extrinsic synthesized blur with the transferred flow map $\tilde{\mathcal{F}}_C$ (right). All flow maps are masked by event masks to explicitly highlight motion-intensive regions.	31
3.9	Ablation study on key hyperparameters. We analyze the impact of (a) the top patches ratio r , (b) the number of extrinsic pairs n , and (c) the orientation threshold ranges $[\tau_{low}, \tau_{high}]$. The best values are marked with red stars.	32
3.10	Visual comparison of blur synthesis under different conditioning. While the baseline using only flow maps (w/o Events) only captures macro-motions, and DA [5] struggles in motion-intensive areas, our FEDA successfully synthesizes blur patterns with high fidelity. The average PSNR scores on the GoPro test set (top) quantitatively show its superior blur synthesis quality.	32
3.11	Illustration of the Motion Source Selector	34
3.12	Additional qualitative results on EBF1, NAFNet, MIMO-UNet. We compare the blurry input (blur), pre-trained baseline (baseline), our adapted model (ours), and ground truth (sharp), with red boxes denoting zoomed-in regions. Our approach consistently restores finer details across both event-based (a) and frame-only (b, c) models.	35
3.13	Additional qualitative examples of dual blur synthesis (intrinsic and extrinsic). Similar to Fig. 3.8 in the main manuscript, each row displays a content source (left), followed by the intrinsic synthesized blur (green box) which utilizes inherent motion. The middle columns show the motion source, and lastly, the resulting extrinsic synthesized blur on the right (blue box).	36
4.1	Overview of our framework. At each denoising step t , the noisy latent z_t^i is processed by the diffusion UNet with Cross-frame Attention, which injects warped Key/Value features from the previous frame z_t^{i-1} to propagate temporal context. The predicted noise ϵ_θ yields an intermediate clean latent $\hat{x}_0^i$, which is refined by Prediction Consistency Guidance toward the previous frame’s fully denoised latent z_0^{i-1} . Separately, Noise Consistency Guidance replaces the stochastic noise with a reliability-weighted blend from the previous frame’s noise η_t^{i-1} . The edited latent $\tilde{x}_0^i$ and noise $\tilde{\eta}_t^i$ are combined via posterior mean sampling to produce z_{t-1}^i	39
4.2	Cross-frame Attention. The Key/Value features of the previous frame $\mathbf{K}_t^{i-1}, \mathbf{V}_t^{i-1}$ are warped using event-based optical flow and blended with the current frame’s features $\mathbf{K}_t^i, \mathbf{V}_t^i$ weighted by the spatial reliability weight r . The resulting blended Keys and Values $\tilde{\mathbf{K}}, \tilde{\mathbf{V}}$ are used in the attention computation alongside the current frame’s Query $\mathbf{Q}_t^i$, propagating temporally consistent structural context into the denoising process.	40

4.3	Noise and Prediction Consistency Guidance. At each denoising step, the UNet takes noisy latent z_t^i and outputs predicted noise ϵ_θ , from which an intermediate clean latent $\hat{x}_0^i$ is estimated. Prediction Consistency Guidance refines $\hat{x}_0^i$ toward the previous frame's clean latent z_0^{i-1} , spatially weighted by the reliability score. Noise Consistency Guidance replaces the stochastic noise with a reliability-weighted blend from the previous frame's noise η_t^{i-1} . The edited $\tilde{x}_0^i$ and $\tilde{\eta}_t^i$ are combined via posterior mean sampling to produce z_{t-1}^i	41
4.4	Qualitative comparison across consecutive frames. MAENet produces overly smoothed results, and DeblurDiff suffers from severe temporal inconsistency. In contrast, our method generates both temporally consistent while preserving its deblurring quality.	43
4.5	Additional qualitative comparison.	43

Chapter 1. Introduction

Visual perception is fundamental to numerous computer vision applications, including autonomous driving and robotics [6]. However, conventional frame-based cameras are fundamentally constrained by their discrete image acquisition process. As illustrated on the left of Fig. 1.1, a standard RGB camera captures a dynamic scene as a series of discrete snapshots separated by a fixed shutter period. During this integration time, any abrupt movement or sudden illumination change results in severe degradations such as motion blur, motion aliasing, and exposure artifacts. Consequently, existing image and video restoration models [7,8] relying solely on these conventional frames often struggle to restore high-quality visuals in highly dynamic and unpredictable real-world environments.

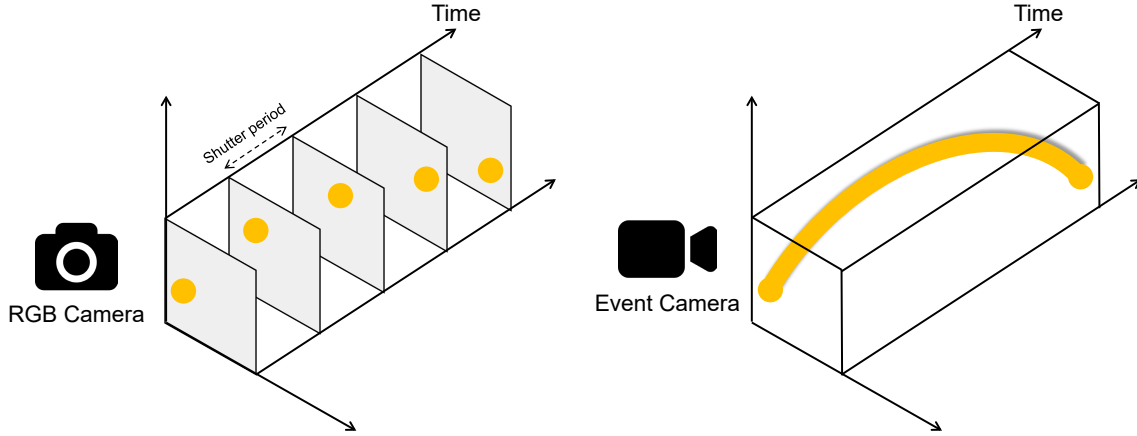


Figure 1.1: **Comparison of visual data acquisition.** A conventional RGB camera captures discrete frames with a fixed shutter period (left), whereas an event camera continuously records the spatiotemporal trajectory of moving objects (right).

To address these limitations, event cameras have emerged as novel vision sensors [9,10]. Unlike conventional cameras that measure absolute intensity synchronously, event cameras respond asynchronously only to pixel-level brightness changes. As shown on the right of Fig. 1.1, this mechanism allows the event camera to record the continuous spatiotemporal trajectory of the moving object without the *blind spots* caused by a shutter period. Benefiting from microsecond-level temporal resolution and a remarkably high dynamic range, event cameras provide rich, continuous intra- and inter-frame cues, opening a new paradigm for overcoming the inherent degradation of frame-based cameras.

Despite the potential of the event modality, utilizing it as a reliable guide for real-world image and video restoration poses inherent challenges. Temporally, it is highly ambiguous to determine the proper amount of events required to restore a specific frame, especially in practical scenarios where the exposure time of the frame-based camera is unknown. Spatially, events are inherently sparse in static or texture-less regions, which limits their effectiveness and reliability as spatial guidance.

Acknowledging these inherent challenges, this dissertation aims to tackle the following core aspects of real-world frame restoration. To ensure that events serve as a robust guide, we carefully integrate techniques to mitigate the aforementioned challenges throughout our studies. Specifically, to address the temporal ambiguity, we leverage frame features to adaptively determine the optimal event usage

for unknown exposures. Furthermore, to handle the spatial sparsity, we apply a simple yet effective masking strategy that restricts event guidance only to regions with active occurrences, thereby preventing unreliable information from degrading the restoration.

The specific real-world aspects addressed in this dissertation, along with our event-guided solutions, are as follows:

- **Blind and Dynamic Exposures:** Real-world cameras continuously adjust their exposure times to accommodate varying lighting conditions. Because existing restoration methods typically assume fixed or predefined exposure times during training, their generalization performance may be limited when encountering unknown and dynamic exposure scenarios.
- **Domain Shift:** Since acquiring large-scale, perfectly aligned real-world deblurring datasets is highly challenging, training deblurring models largely relies on synthetic data. However, the domain shift between synthetic and real-world motions can restrict the restoration quality when these models are applied to practical scenarios.
- **Computational Feasibility in Generative Models:** The massive spatial-temporal attention costs of Video Diffusion Models often restrict their practical deployment. As a more feasible alternative, pre-trained Image Diffusion Models (IDMs) are frequently adopted to achieve high-quality spatial restoration without the training burden of VDMs. Nevertheless, this frame-by-frame generation process inevitably suffers from temporal inconsistency.

Building upon this motivation, this dissertation proposes three event-guided frameworks that achieve robust image and video restoration by carefully managing the inherent limitations of event cameras. The primary contributions of this dissertation are threefold:

- **Chapter 2. *Learning Event-guided Exposure-agnostic Video Frame Interpolation via Adaptive Feature Blending***¹: We introduce an exposure-agnostic framework for joint video frame interpolation and motion deblurring. To overcome the **temporal ambiguity** of unknown exposures, we implicitly model temporal constraints from continuous event streams and adaptively blend cross-modal features, ensuring robust generalization across dynamic lighting conditions.
- **Chapter 3. *EvBS: Event-guided Blur Synthesis for Domain-adaptive Motion Deblurring***²: We present an event-guided blur synthesis pipeline to mitigate the **domain shift** problem. By exploiting the spatiotemporal precision of events, we explicitly decouple motion from spatial content. This allows for the synthesis of realistic training pairs with diverse motions, effectively calibrating pre-trained models to target real-world distributions.
- **Chapter 4. *EvZVD: Event-guided Zero-shot Video Deblurring Using Pre-trained Image Deblurring Diffusion Models***: We propose a training-free, zero-shot framework that enhances the temporal consistency of pre-trained image diffusion models. To prevent artifacts caused by the **spatial sparsity** of events, we employ event density as a reliability metric. This adaptively modulates the generative process, achieving a balance between high-fidelity spatial restoration and temporal stability without massive computational costs.

Finally, **Chapter 5** concludes this dissertation with a summary of the core findings and a brief discussion of future research directions toward robust neuromorphic vision.

¹This chapter is based on the work published in the British Machine Vision Conference (BMVC) 2025.

²This chapter is based on the work currently under review at ACM Multimedia (ACM MM) 2026.

Chapter 2. Learning Event-guided Exposure-agnostic Video Frame Interpolation via Adaptive Feature Blending

2.1 Introduction

Video frame interpolation (VFI) aims to generate intermediate frames between consecutive inputs, converting low-frame-rate videos into high-frame-rate ones [7]. Recent advances in event cameras [9], which asynchronously capture per-pixel brightness changes with high temporal resolution, have significantly improved VFI performance [11–16], particularly in exposure-specific settings where the camera exposure time is fixed and known. However, in real-world scenarios, exposure time often varies and is blind due to auto-exposure mechanisms [17], resulting in inputs with inconsistent motion blur. This variability poses challenges for conventional VFI models, which are typically designed for either consistently sharp or blurry inputs. Consequently, there is a growing need for exposure-agnostic VFI, which can handle unknown and dynamic exposure conditions and is thus essential for practicality.

Recently, a few event-guided methods [1, 18] have been proposed for exposure-agnostic video restoration. UEVD [18] introduced an event-guided method for video deblurring under blind exposure conditions. However, it is limited to restoring a single sharp frame within the exposure window corresponding to the blurry input, and requires a separate VFI step to generate high-frame-rate outputs. EBFI [1] later proposed a unified framework that restores arbitrary target frames under blind exposure from a single blurry frame by modulating events (see Fig. 2.1a). Yet, the lack of temporal constraints — that is, leveraging the temporal relationships between neighboring frames for consistent and accurate guidance — leads to performance degradation, especially when input videos are captured at very low frame rates.

To address this limitation, a promising direction is to introduce *temporal constraints across consecutive frames* since frames closer to the target timestamp often offer more relevant cues. Flow-based methods [11–13, 19, 20] offer a natural way to model such constraints by aligning frame-to-frame correspondences. However, their effectiveness is limited under blind exposure, where severe motion blur degrades both frame quality and flow estimation [7, 9]. An alternative is to leverage the *mutual relationship* between frames and events [21], which introduces two key challenges: (1) how to sample events between the target timestamp and the unknown exposure period in a way that reflects the mutual relationship between frames and events, and (2) how to impose temporal constraints across consecutive features while jointly considering temporal proximity and spatial relevance.

In this paper, we address the aforementioned challenges through a simple yet effective adaptive feature blending strategy (see Fig. 2.1b). To this end, we introduce a novel learning framework that consists of two key modules: the Target-adaptive Event Sampling (TES) and the Target-adaptive Importance Mapping (TIM). TES samples events around the target timestamp and the unknown exposure period, which are fused with the corresponding blurry frames to guide target sharp frame synthesis. TIM is designed to generate an importance map that adaptively weights features from consecutive frames based on their temporal proximity and spatial relevance to the target. This allows our model to incorporate temporal constraints and maintain performance even as the target timestamp deviates (see Fig. 2.1c). Experiments on synthetic and real-world datasets show that our method consistently outperforms existing approaches under blind exposure, achieving at least a 2dB PSNR improvement.

In summary, our contributions are threefold: 1) We propose a simple yet effective feature blending

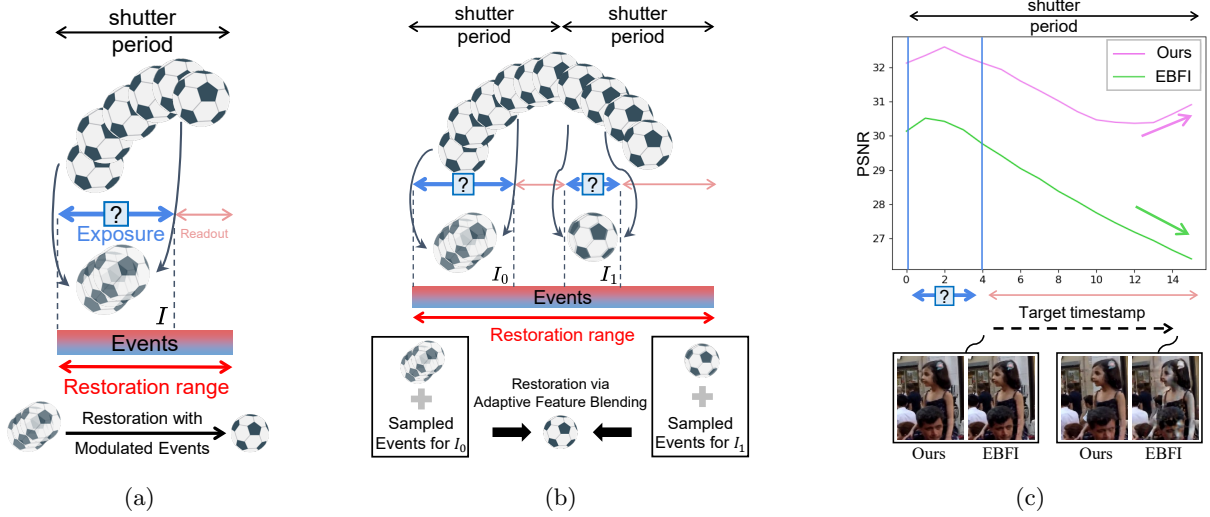


Figure 2.1: **Schematic comparison with a prior event-guided exposure-agnostic VFI method.** (a) EBFI [1] restores sharp frames from a single blurry input via event modulation, but degrades as the target timestamp deviates due to the lack of temporal constraints. (b) Our method addresses this by explicitly leveraging temporal constraints via adaptive feature blending. (c) PSNR comparison on GoPro [2] shows that our method maintains more stable performance across varying timestamps.

strategy that, for the first time, incorporates temporal constraints into event-guided VFI under unknown and dynamic exposure conditions. 2) We design a novel unified framework with two synergistic modules: TES, which samples events around the target timestamp and unknown exposure; and TIM, which generates an importance map to adaptively weight features based on temporal and spatial relevance. 3) We validate our approach through extensive experiments on synthetic and real-world datasets, showing its effectiveness across diverse blind exposure scenarios.

2.2 Related Work

Exposure-specific Video Restoration. Most prior work assumes a known and fixed exposure time when restoring sharp, high-frame-rate videos. For video frame interpolation (VFI), existing methods are broadly categorized into flow-based [19, 20, 22–25], kernel-based [26, 27], architecture-based [28–31], and event-guided [11–13] approaches. Motion deblurring has also been studied independently to restore one sharp frame from each blurry input [21, 32–37]. Rather than treating VFI and motion deblurring as separate tasks, recent studies have explored unified frameworks that address both jointly, using either RGB-only [38, 39] or event-guided settings [14–16]. However, their reliance on fixed exposure limits their applicability to real-world scenarios where exposure is unknown and dynamically changing.

Exposure-agnostic Video Restoration. A few RGB-only methods [40, 41] have been proposed for video restoration under unknown exposure conditions. However, due to the lack of precise motion cues, these approaches suffer from motion ambiguity [1, 16], making it difficult to accurately recover sharp frames. To overcome this, event cameras offer a promising alternative by providing temporally dense and precise motion information. Building on this idea, Kim *et al.* [18] proposed an event-guided motion deblurring method under blind exposure. However, their approach is limited to reconstructing a single sharp frame within the exposure window and requires an additional VFI step to generate high-frame-rate

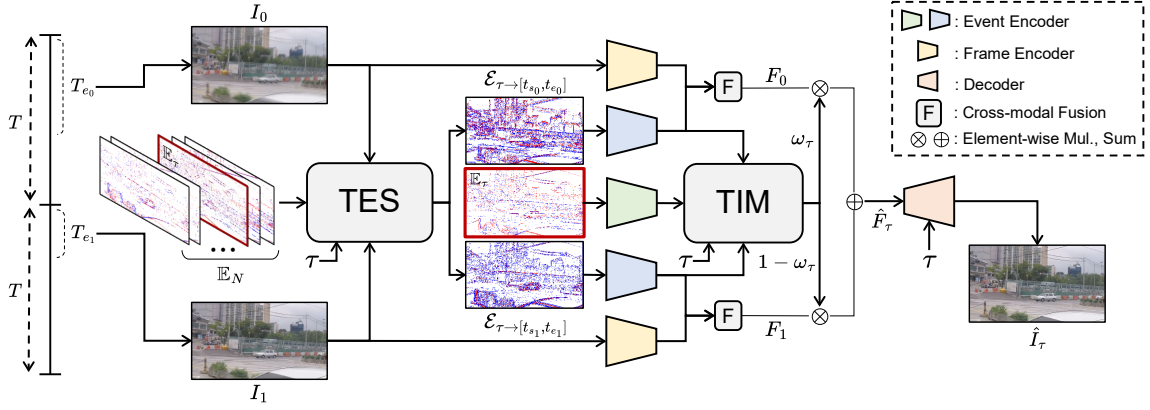


Figure 2.2: **Overview of our framework.** Given blurry frames (I_0, I_1) with unknown exposures (T_{e_0}, T_{e_1}) , stacked events $\mathbb{E}_N$ over $2T$, and target timestamp τ , the model reconstructs the target frame $\hat{I}_\tau$. TES samples events around τ and unknown exposure, which are fused with frames. TIM generates an importance map ω_τ to adaptively blend the fused features (F_0, F_1) into F_τ , which is then decoded into $\hat{I}_\tau$. Shared encoders are color-coded.

output. Weng *et al.* [1] presented a framework for interpolating arbitrary frames by modulating a blurry frame with events, but the lack of temporal constraints leads to degraded performance under severe blur and sparse temporal sampling. We address this limitation with a unified framework that explicitly incorporates temporal constraints into event-guided, exposure-agnostic VFI.

2.3 Method

In this section, we first formulate the problem based on the mutual relationship between frames and events using the Event-based Double Integral (EDI) model [21]. We then present an overview of our framework (Sec.2.3.1), followed by details of the proposed TES and TIM modules (Sec.2.3.2, 2.3.3).

Problem Formulation. Event cameras capture events, denoted as $e(\mathcal{T})$, at timestamp $\mathcal{T}$ across the pixel space asynchronously [9]. These events are recorded when the logarithmic change in intensity exceeds a contrast threshold c . During the exposure time T_e , the captured frame I is derived by averaging latent frames $L(t)$ present in T_e as:

$$I = \frac{1}{T_e} \int_{t_s}^{t_e} L(t) dt, \quad (2.1)$$

where L , t_s , and t_e denote the latent frame, the starting timestamp of T_e , and the end timestamp of T_e , respectively.

Given latent frames $L(\tau)$ and $L(t)$ at different timestamps τ and t during the exposure time, and the events spanning the interval $[\tau, t]$, we can establish the following relationship between the latent frames:

$$L(t) = L(\tau) \exp(c \cdot \mathcal{E}_{\tau \rightarrow t}), \quad (2.2)$$

where $\mathcal{E}_{\tau \rightarrow t} = \int_\tau^t e(x) dx$ denotes a 2D tensor obtained by stacking events from τ to t . Combining Eqs. (2.1) and (2.2), the following Event-based Double Integral (EDI) model [21] can be derived:

$$L(\tau) = \frac{T_e \cdot I}{\int_{t_s}^{t_e} \exp(c \cdot \mathcal{E}_{\tau \rightarrow t}) dt}. \quad (2.3)$$

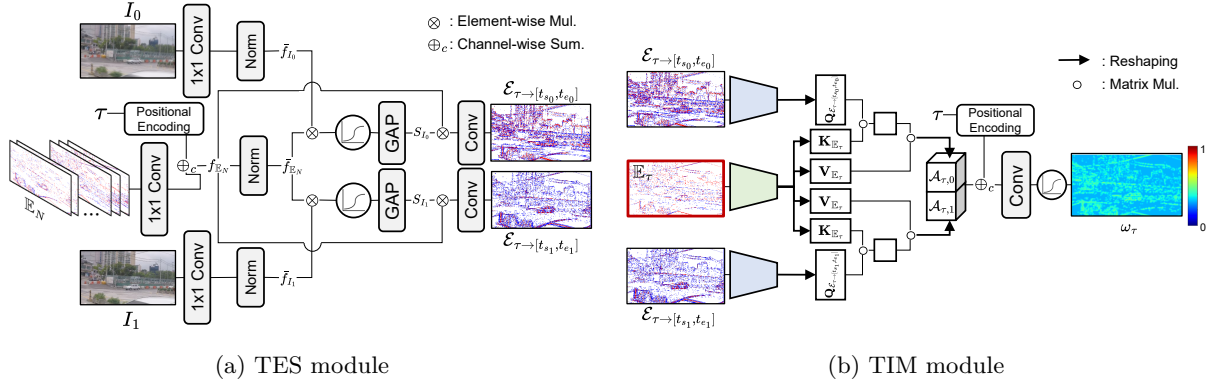


Figure 2.3: Architectures of the TES (Sec. 2.3.2) and TIM (Sec. 2.3.3) modules.

The EDI model describes how the latent frame $L(\tau)$ can be reconstructed by combining the captured frame I with stacked events over the exposure duration T_e . In VFI scenarios where the target timestamp τ lies beyond T_e , EVDI [15] extended such mutual relationship by using events from τ to T_e to restore sharp frames outside the exposure window. However, this approach requires precise knowledge of T_e , limiting its practicality. Nonetheless, this suggests that if we can approximate events centered around τ and T_e without knowing T_e , it is still possible to recover $L(\tau)$ based on Eq. (2.3). We denote this event set as $\mathcal{E}_{\tau \rightarrow [t_s, t_e]}$.

2.3.1 Overview

We adopt a UNet-based encoder-decoder model [42], comprising three encoders and two decoders, with a refinement block [15] appended at the final decoder stage. The model takes as input two consecutive captured frames (I_0, I_1) — each recorded under unknown exposure durations (T_{e_0}, T_{e_1}) — along with the corresponding events recorded during their shutter periods $2T$. These events are converted into event frames [43] by pixel-wise accumulation over N bins, yielding $\mathbb{E}_N \in \mathbb{R}^{N \times H \times W}$, where N denotes the number of temporal slices.

To apply Eq. (2.3), we sample a subset of events $\mathcal{E}_{\tau \rightarrow [t_s, t_e]}$ from the stacked input $\mathbb{E}_N$, centered around the target timestamp τ and unknown exposure T_e . To this end, we introduce the Target-adaptive Event Sampling (TES) module (Sec. 2.3.2), samples $(\mathcal{E}_{\tau \rightarrow [t_{s_0}, t_{e_0}]}, \mathcal{E}_{\tau \rightarrow [t_{s_1}, t_{e_1}]})$ accordingly. Frame and event features from (I_0, I_1) and sampled events are fused via a cross-modal technique [35] to produce (F_0, F_1) .

To leverage temporal constraints, we introduce the Target-adaptive Importance Mapping (TIM) module (Sec. 2.3.3), which generates an importance map ω_τ based on the temporal proximity and spatial relevance of (F_0, F_1) to τ . This map guides adaptive blending: temporally closer features are emphasized, while spatially relevant cues complement them. The resulting fused feature $\hat{F}_\tau$ is decoded to reconstruct the target frame $\hat{I}_\tau$.

2.3.2 Target-adaptive Event Sampling

Recalling Eq. (2.3), the frames (I_0, I_1) should be combined with events centered around the target timestamp τ and the unknown exposure durations (T_{e_0}, T_{e_1}). To achieve this, we follow prior methods [1, 18] by using the captured frames as references under blind exposure. However, unlike previous approaches that focus solely on the unknown exposure period, our module is additionally designed to account for

the target timestamp τ .

A detailed illustration of the TES module is shown in Fig. 2.3a. For clarity, we describe the process of sampling the stacked events for I_0 , denoted as $\mathcal{E}_{\tau \rightarrow [t_{s_0}, t_{e_0}]}$. Given $\mathbb{E}_N$, I_0 , and τ as inputs, the TES module is formulated as:

$$\mathcal{E}_{\tau \rightarrow [t_{s_0}, t_{e_0}]} = \text{TES}(\mathbb{E}_N, I_0, \tau). \quad (2.4)$$

We first extract features from $\mathbb{E}_N$ and I_0 . To incorporate the target timestamp τ , we add its positional encoding to the event features channel-wise, yielding $f_{\mathbb{E}_N}$. We then compute the correlation score $\mathcal{S}_{I_0}$ between the normalized features of $\bar{f}_{I_0}$ and $\bar{f}_{\mathbb{E}_N}$ as:

$$\mathcal{S}_{I_0} = \text{GAP}(\sigma(\bar{f}_{I_0} \otimes \bar{f}_{\mathbb{E}_N})), \quad (2.5)$$

where $\otimes$, σ , and GAP denote element-wise multiplication, the sigmoid function, and global average pooling, respectively. Using $\mathcal{S}_{I_0}$, we perform element-wise multiplication with $f_{\mathbb{E}_N}$ to sample the event features centered around τ and T_{e_0} . A convolution is then applied to produce the final representation $\mathcal{E}_{\tau \rightarrow [t_{s_0}, t_{e_0}]}$. The counterpart $\mathcal{E}_{\tau \rightarrow [t_{s_1}, t_{e_1}]}$ is derived analogously using I_1 , and both event representations are fused with (I_0, I_1) using the fusion method from [35] (denoted as ‘F’ in Fig. 2.2).

2.3.3 Target-adaptive Importance Mapping

Although TES provides relevant events, restoration degrades as τ shifts from the exposure due to event sensor noise [9]. When τ nears T_{e_1} , relying only on I_0 becomes suboptimal, making features from I_1 more informative and underscoring the need for temporal constraints.

To address this, we introduce the TIM module, which estimates an importance map ω_τ to adaptively weight fused features (F_0, F_1) based on their temporal proximity and spatial relevance to the target. Given the sampled events $(\mathcal{E}_{\tau \rightarrow [t_{s_0}, t_{e_0}]}, \mathcal{E}_{\tau \rightarrow [t_{s_1}, t_{e_1}]})$, the τ -centered event stack $\mathbb{E}_\tau$, and the timestamp τ , TIM estimates ω_τ as:

$$\omega_\tau = \text{TIM}(\mathcal{E}_{\tau \rightarrow [t_{s_0}, t_{e_0}]}, \mathcal{E}_{\tau \rightarrow [t_{s_1}, t_{e_1}]}, \mathbb{E}_\tau, \tau). \quad (2.6)$$

A detailed illustration of the TIM module is shown in Fig. 2.3b. We first extract features from the sampled events $(\mathcal{E}_{\tau \rightarrow [t_{s_0}, t_{e_0}]}, \mathcal{E}_{\tau \rightarrow [t_{s_1}, t_{e_1}]})$ and the τ -centered event stack $\mathbb{E}_\tau$. To estimate spatial relevance efficiently, we apply channel-wise attention [44]: features from the sampled events are used as queries, and those from $\mathbb{E}_\tau$ serve as keys and values. After flattening the spatial dimensions, the attended feature $\mathcal{A}_{\tau,0}$ is computed as:

$$\mathcal{A}_{\tau,0} = \mathbf{V}_{\mathbb{E}_\tau} \circ \text{softmax} \left(\frac{\mathbf{Q}_{\mathcal{E}_{\tau \rightarrow [t_{s_0}, t_{e_0}]}}^T \circ \mathbf{K}_{\mathbb{E}_\tau}}{\sqrt{d_k}} \right), \quad (2.7)$$

where d_k denotes the key dimension. $\mathcal{A}_{\tau,1}$ is computed analogously. The two attended features are concatenated and combined with the positional encoding of τ via channel-wise addition to encode temporal information. The result is passed through convolutional layers and a sigmoid activation to produce the importance map ω_τ , which encodes both temporal proximity and spatial relevance.

The target feature $\hat{F}_\tau$ is obtained via an adaptive blending strategy that applies temporal constraints:

$$\hat{F}_\tau = (\omega_\tau \otimes F_0) \oplus ((1 - \omega_\tau) \otimes F_1). \quad (2.8)$$

A decoder then transforms $\hat{F}_\tau$ into the target sharp frame $\hat{I}_\tau$. This blending strategy enables the model to effectively leverage temporal constraints between consecutive features. The entire framework is trained end-to-end using the Charbonnier loss [45]:

$$\mathcal{L} = \sqrt{\|I_\tau^{gt} - \hat{I}_\tau\|^2 + \epsilon^2}, \text{ with } \epsilon = 1\text{e-}6. \quad (2.9)$$

Table 2.1: Summary of feature comparisons among different video deblurring methods. Our method simultaneously satisfies all the key properties.

Method	Frame	Events	Agnostic	Unified	T-Constraint
NAFNet [37] + RIFE [19]	✓	✗	✗	✗	✓(via RIFE)
UTI [40]	✓	✗	✓	✓	✓
VIDUE [41]	✓	✗	✓	✓	✓
UEVD [18] + TL [11]	✓	✓	✓(via UEVD)	✗	✓(via TL)
EVDI [15]	✓	✓	✗	✓	✓
REFID [16]	✓	✓	✗	✓	✓
EBFI [1]	✓	✓	✓	✓	✗
Ours	✓	✓	✓	✓	✓

2.4 Experiments

2.4.1 Experimental Setups

Datasets. We evaluate on GoPro [2] (with synthetic events from ESIM [46]), HighREV [16] (real events, 1632×1224), and RealBlur-DAVIS [1] (real blurry videos and events). To simulate motion blur in GoPro and HighREV, we average m consecutive frames within the exposure duration and downsample them to shutter periods of 10 and 16 frames, denoted as $10\downarrow$ and $16\downarrow$.

Implementation Details. Our framework is trained end-to-end without pre-training. We train on GoPro for 300 epochs using AdamW [47] with a learning rate of $5e-4$, scheduled by cosine annealing [48], and fine-tune on HighREV for 20 epochs with a learning rate of $1e-4$. We follow the official data splits. During training, m is randomly sampled within the shutter period (e.g., $m \in [1, 9]$ for $10\downarrow$), and inputs are cropped to 128×128 (GoPro) or 256×256 (HighREV). Target timestamps are randomly selected from available indices.

Evaluation Protocols. We evaluate under two settings: (1) *Symmetric exposure* with fixed m values ($[1, 5, 9]$ for $10\downarrow$, $[1, 5, 11, 15]$ for $16\downarrow$), and (2) *RandEx*, where m is randomly sampled as in training. Metrics include PSNR (dB), SSIM [49], and LPIPS [50]. We additionally report temporal consistency metrics (tOF, tLP [51]) and NIQE [52] for RealBlur-DAVIS, which lacks ground-truth frames.

2.4.2 Comparison to the State-of-the-arts

Table 2.1 shows the summary of each compared method. All methods are retrained for fair comparison, except TimeLens (TL), which is evaluated using its pre-trained model due to unavailable training code. Exposure-specific baselines are provided with ground-truth exposure information. All results are averaged across the restored frames.

Quantitative Evaluation. Tables 2.2 and 2.3 show results on GoPro and HighREV under $10\downarrow$ and $16\downarrow$ settings. As expected, RGB-only and cascaded methods tend to yield lower performance, likely due to motion ambiguity and error propagation, respectively. Exposure-specific models (REFID, EVDI), trained with known exposure times, benefit from temporal alignment and produce comparable results. EBFI, despite being exposure-agnostic, performs competitively on GoPro, but its performance declines on HighREV, where larger pixel displacement highlights the importance of temporal constraints. Our method achieves the highest PSNR, SSIM, and LPIPS on both GoPro and HighREV, with margins up to 2.85dB over the second-best. Temporal consistency (tOF, tLP) and NIQE results on RealBlur-DAVIS

Table 2.2: Quantitative results on GoPro and HighREV (10↓). **Bold** and underlined indicate the best and second-best scores, respectively.

Method	PSNR↑ / SSIM↑ / LPIPS↓			
	9+1	5+5	1+9	RandEx
<i>GoPro-10↓ Dataset</i>				
NAFNet [37]+RIFE [19]	23.15 / 0.736 / 0.164	21.82 / 0.665 / 0.188	18.78 / 0.520 / 0.282	21.05 / 0.633 / 0.219
UTI [40]	23.84 / 0.805 / 0.205	23.25 / 0.771 / 0.166	20.95 / 0.640 / 0.170	22.17 / 0.714 / 0.189
VIDUE [41]	25.37 / 0.818 / 0.181	26.24 / 0.835 / 0.145	24.76 / 0.764 / 0.146	25.86 / 0.819 / 0.152
UEVD [18]+TL [11]	23.40 / 0.720 / 0.194	23.31 / 0.755 / 0.161	26.33 / 0.866 / 0.103	23.51 / 0.757 / 0.164
EVDI [15]	28.18 / 0.898 / <u>0.052</u>	27.73 / 0.884 / <u>0.067</u>	26.07 / 0.834 / 0.084	27.51 / 0.877 / <u>0.071</u>
REFID [16]	<u>31.08</u> / 0.939 / 0.077	<u>31.28</u> / <u>0.941</u> / 0.071	30.27 / <u>0.927</u> / <u>0.078</u>	<u>31.03</u> / <u>0.938</u> / 0.074
EBFI [1]	31.06 / <u>0.942</u> / 0.072	31.08 / 0.940 / 0.071	<u>30.32</u> / 0.921 / 0.087	30.89 / 0.936 / 0.075
Ours	33.22 / 0.960 / 0.050	33.61 / 0.963 / 0.042	32.87 / 0.954 / 0.048	33.39 / 0.961 / 0.045
<i>HighREV-10↓ Dataset</i>				
NAFNet [37]+RIFE [19]	25.09 / 0.805 / 0.446	24.90 / 0.874 / 0.401	26.07 / 0.897 / 0.361	25.20 / 0.844 / 0.402
UTI [40]	26.22 / 0.834 / 0.371	26.15 / 0.821 / 0.395	26.51 / 0.851 / 0.395	26.19 / 0.831 / 0.398
VIDUE [41]	26.65 / 0.847 / 0.375	27.43 / 0.860 / 0.367	25.31 / 0.841 / 0.387	26.76 / 0.850 / 0.373
UEVD [18]+TL [11]	26.02 / 0.844 / 0.399	27.41 / 0.854 / 0.378	27.77 / 0.861 / 0.337	27.43 / 0.854 / 0.373
EVDI [15]	30.26 / 0.910 / <u>0.278</u>	29.75 / 0.896 / <u>0.277</u>	27.01 / 0.839 / 0.292	29.25 / 0.887 / <u>0.273</u>
REFID [16]	<u>33.60</u> / <u>0.937</u> / 0.306	<u>34.18</u> / <u>0.938</u> / 0.296	<u>33.38</u> / <u>0.928</u> / <u>0.260</u>	<u>33.91</u> / <u>0.936</u> / 0.292
EBFI [1]	28.23 / 0.898 / 0.362	27.84 / 0.887 / 0.342	26.25 / 0.855 / 0.279	27.55 / 0.882 / 0.333
Ours	35.73 / 0.948 / 0.263	36.27 / 0.949 / 0.250	35.45 / 0.942 / 0.248	36.02 / 0.947 / 0.253

are provided in the supplementary material (Sec. 2.6.1).

Qualitative Evaluation. We present qualitative results on RealBlur-DAVIS (real captured video) in Fig. 2.4, where our method produces temporally consistent frames with minimal artifacts. Additional comparisons on GoPro, HighREV (with synthetic events), and RealBlur-DAVIS, along with supplementary video clips demonstrating qualitative results and arbitrary temporal upscaling, are provided in supplementary material (Sec. 2.6.2).

2.4.3 Analysis

Analysis of the TES Module. We analyze whether TES effectively samples events around the target timestamp τ and the unknown exposure T_e . To this end, we apply a gradient-based interpretation [53] and compute importance scores as $\frac{\partial \mathcal{E}_{\tau \rightarrow T_e}}{\partial \mathbf{E}_N}$, indicating the contribution of each event to the sampled output. Given that input events are divided into N temporal bins and stacked channel-wise, we report the average importance score per channel on GoPro-10↓. As shown in Fig. 2.5, TES consistently focuses on events near both T_e and τ , even when τ lies outside the exposure window.

We present ablation results on GoPro-10↓ (RandEx) in the ‘TES’ section of Table 2.4. We evaluate the impact of removing key components from the TES module: (1) the sampling process in Eq. (2.5) and (2) the positional encoding (PE) of the target timestamp τ . Removing the sampling process hinders the TES module from focusing on events near τ and the blind exposure T_e , while excluding PE causes the module to ignore τ and focus only on T_e . These degradations confirm that both components are essential for effective event sampling. Further qualitative ablations are included in the supplementary (Sec. 2.6.3).

Analysis of the TIM Module. We visualize the importance map ω_τ in Fig. 2.6 using GoPro-16↓ data with $m = 7$, where the exposure ranges are $[1, 7]$ for I_0 and $[17, 23]$ for I_1 . As τ approaches the exposure of I_0 , ω_τ assigns greater weights to features from I_0 , and likewise shifts toward I_1 as τ nears its

Table 2.3: Quantitative results on GoPro and HighREV (16 $\downarrow$). **Bold** and underlined indicate the best and second-best scores, respectively.

Method	PSNR $\uparrow$ / SSIM $\uparrow$ / LPIPS $\downarrow$				
	15+1	11+5	5+11	1+15	RandEx
<i>GoPro-16$\downarrow$ Dataset</i>					
NAFNet [37]+RIFE [19]	21.37 / 0.645 / 0.297	21.10 / 0.627 / 0.284	19.75 / 0.551 / 0.270	17.32 / 0.430 / 0.322	20.03 / 0.569 / 0.233
UTI [40]	21.43 / 0.672 / 0.286	21.19 / 0.659 / 0.273	20.23 / 0.591 / 0.272	19.64 / 0.534 / 0.298	20.74 / 0.609 / 0.234
VIDUE [41]	22.44 / 0.686 / 0.298	23.27 / 0.721 / 0.270	23.10 / 0.697 / 0.231	21.58 / 0.623 / 0.220	22.89 / 0.695 / 0.253
UEVD [18]+TL [11]	17.04 / 0.316 / 0.472	17.26 / 0.334 / 0.459	17.92 / 0.376 / 0.444	18.41 / 0.401 / 0.438	17.64 / 0.357 / 0.453
EVDI [15]	27.38 / 0.881 / <u>0.083</u>	27.16 / 0.874 / <u>0.084</u>	26.36 / 0.847 / <u>0.092</u>	24.15 / 0.774 / <u>0.115</u>	26.52 / 0.853 / <u>0.088</u>
REFID [16]	<u>28.98</u> / 0.912 / 0.118	<u>29.11</u> / 0.913 / 0.115	<u>28.84</u> / <u>0.909</u> / 0.115	<u>28.16</u> / <u>0.896</u> / 0.123	<u>28.87</u> / <u>0.909</u> / 0.116
EBFI [1]	28.78 / <u>0.913</u> / 0.108	28.92 / <u>0.914</u> / 0.106	28.48 / 0.903 / 0.116	27.64 / 0.883 / 0.134	28.59 / 0.906 / 0.113
Ours	30.97 / 0.937 / 0.081	31.39 / 0.941 / 0.074	31.18 / 0.938 / 0.075	30.07 / 0.925 / 0.085	31.14 / 0.938 / 0.076
<i>HighREV-16$\downarrow$ Dataset</i>					
NAFNet [37]+RIFE [19]	19.85 / 0.336 / 0.744	19.70 / 0.285 / 0.770	19.91 / 0.312 / 0.764	20.35 / 0.332 / 0.734	19.53 / 0.287 / 0.765
UTI [40]	21.02 / 0.333 / 0.717	21.83 / 0.369 / 0.685	20.68 / 0.350 / 0.756	20.74 / 0.359 / 0.743	21.16 / 0.347 / 0.712
VIDUE [41]	21.63 / 0.356 / 0.721	21.82 / 0.364 / 0.718	21.63 / 0.370 / 0.713	21.21 / 0.372 / 0.704	21.64 / 0.366 / 0.715
UEVD [18]+TL [11]	24.26 / 0.813 / 0.433	24.94 / 0.818 / 0.425	26.13 / 0.827 / 0.407	26.92 / 0.828 / 0.379	25.58 / 0.822 / 0.414
EVDI [15]	28.75 / 0.893 / <u>0.324</u>	28.62 / 0.887 / <u>0.323</u>	27.09 / 0.870 / <u>0.339</u>	24.16 / 0.786 / 0.386	27.33 / 0.875 / <u>0.336</u>
REFID [16]	<u>30.51</u> / <u>0.921</u> / 0.349	<u>30.91</u> / <u>0.922</u> / 0.346	<u>30.68</u> / <u>0.918</u> / 0.346	<u>29.59</u> / <u>0.907</u> / <u>0.342</u>	<u>30.60</u> / <u>0.919</u> / 0.346
EBFI [1]	26.28 / 0.883 / 0.380	26.18 / 0.878 / 0.370	25.24 / 0.860 / 0.350	23.95 / 0.829 / 0.382	25.54 / 0.866 / 0.355
Ours	33.36 / 0.937 / 0.312	33.93 / 0.938 / 0.307	33.85 / 0.937 / 0.303	32.70 / 0.929 / 0.282	33.69 / 0.937 / 0.303

exposure. These results show that TIM adaptively emphasizes features that are temporally closer and spatially relevant to the target timestamp.

We report ablation results for the TIM module in the ‘TIM’ section of Table 2.4. Removing the target timestamp τ causes convergence failure, underscoring its importance for generating ω_τ . Disabling the attention mechanism in Eq. 2.7 also degrades performance, confirming the benefit of spatial relevance. We also evaluate two variants: replacing ω_τ with a constant value (0.5) and swapping its order during feature blending in Eq. (2.8). Both degrade performance, with the swapped version performing notably worse, validating the importance and learned structure of ω_τ . Qualitative ablations are provided in the supplementary (Sec. 2.6.3).

2.5 Conclusion

In this work, we proposed a simple yet effective framework for event-guided, exposure-agnostic video frame interpolation that explicitly incorporates temporal constraints via adaptive feature blending. Our method integrates two synergistic modules: the Target-adaptive Event Sampling module for event sampling around the target timestamp and unknown exposure, and the Target-adaptive Importance Mapping module for weighting features based on temporal and spatial relevance. Extensive experiments validate its effectiveness across synthetic and real-world datasets.

In this supplementary material, we first present additional results in Sec. 2.6. Next, we provide network details in Sec. 2.6.4. Finally, Sec. 2.6.5 discusses the limitations of our approach.



Figure 2.4: Qualitative results on RealBlur-DAVIS (real blur & events)

2.6 Appendix

2.6.1 More Quantitative Results

We additionally report temporal consistency metrics (tOF, tLP [51]) and the no-reference quality metric NIQE [52] for RealBlur-DAVIS, which lacks ground-truth frames. As shown in Table 2.5, our method achieves the most temporally coherent results on both GoPro and HighREV datasets. Furthermore, Table 2.6 shows that our method also achieves the best NIQE scores, indicating superior perceptual quality on real-world blurry videos.

2.6.2 More Qualitative Results

We present additional visual comparisons on GoPro, HighREV, and RealBlur-DAVIS in Fig.2.10, Fig.2.11, Fig.2.12, and Fig.2.13. On the synthetic GoPro dataset (Fig.2.10), our method consistently produces the most faithful results across diverse scenes. For HighREV (Fig.2.11), visualizations across various target timestamps show that our method maintains high-quality reconstructions with minimal artifacts.

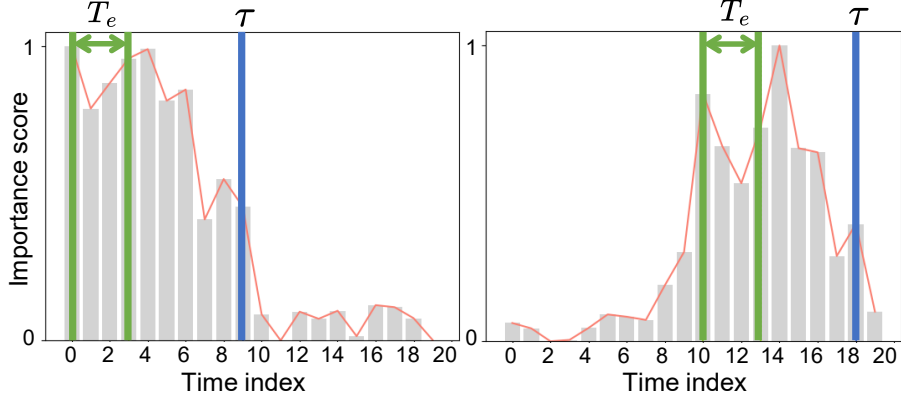


Figure 2.5: Channel-wise importance score $\frac{\partial \mathcal{E}_{\tau \rightarrow [t_s, t_e]}}{\partial \mathbb{E}_N}$ indicating the sampling contribution of each time index. TES effectively selects events near the unknown exposure T_e (green) and target timestamp τ (blue), regardless of whether τ lies outside the exposure window.

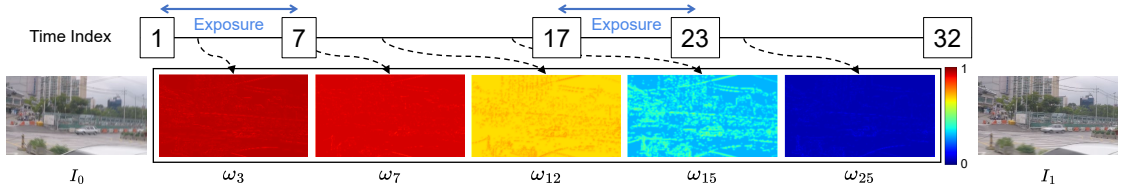


Figure 2.6: Visualization of importance maps ω_τ for $\tau \in \{3, 7, 12, 15, 25\}$, with symmetric exposure durations ($m = 7$) for I_0 and I_1 .

On the RealBlur-DAVIS dataset (Figs.2.12 and 2.13), two-stage methods (UEVD + TL) yield the least effective results, while event-guided approaches (Ours, EBFI, REFID) outperform the RGB-only baseline (VIDUE) by leveraging precise motion cues from events. Notably, as seen in Fig. 2.12, event-guided methods can even recover objects missing from RGB frames thanks to the high dynamic range of event sensors. Among them, our method delivers the most reliable results across all target timestamps.

We also provide two types of supplementary MP4 video clips: (1) comparisons with strong baselines (EBFI [1] and REFID [16]), and (2) demonstrations showing our method upscaling low-frame-rate blurry inputs into sharp, high-frame-rate outputs at arbitrary temporal scales. These videos further highlight the temporal coherence and scalability of our approach.

2.6.3 More Analysis

We qualitatively ablate the TES module by removing the sampling step and the positional encoding of the target timestamp τ . Figure 2.7 illustrates the resulting changes in channel-wise importance scores. Without sampling (Fig.2.7(a)), TES fails to focus on events around τ and T_e ; without τ (Fig.2.7(b)), it only attends to T_e . These results highlight the critical role of both components in effective event sampling.

To further evaluate the effectiveness of the TIM module, Fig. 2.8 compares the PSNR over time between our full model and a variant that uses only a single frame and its corresponding events. The results underscore the role of TIM in enhancing temporal consistency across frames.

Computational Cost. Table 2.7 compares model parameters, FLOPs, and runtime. EVDI [15] has the lowest cost due to its lightweight design, but shows limited performance under unknown and dynamic

Table 2.4: Ablation study on GoPro-10↓ (RandEx). All ablations are retrained from scratch, except Swap ω_τ .

Ablation	PSNR↑	SSIM↑	LPIPS↓
TES (Fig. 2.3a)			
Remove Sampling	32.05	0.948	0.058
Remove τ	31.12	0.937	0.073
TIM (Fig. 2.3b)			
Remove τ		×	
Remove Eq. 2.7	32.24	0.945	0.051
Fixed $\omega_\tau = 0.5$	31.51	0.939	0.068
Swap ω_τ	20.43	0.588	0.194
Ours	33.39	0.961	0.045

Table 2.5: tOF↓ / tLP↓ evaluations on GoPro and HighREV with RandEX setting. **Bold** indicates the best scores.

Dataset	tOF↓ / tLP↓							
	Ours	EBFI	REFID	EVDI	UEVD+TL	VIDUE	UTI	NAFNet+RIFE
GoPro-10↓	0.0067 / 0.6366	0.0082 / 0.8873	0.0078 / 0.7618	0.0104 / 0.8324	0.0206 / 1.4478	0.0232 / 1.5676	0.0410 / 3.9124	0.0495 / 4.6272
GoPro-16↓	0.0867 / 0.7618	0.1064 / 0.9227	0.1073 / 0.9827	0.1270 / 1.0581	0.4690 / 6.2587	0.3370 / 2.8956	0.3980 / 4.9081	0.4040 / 5.0532
HighREV-10↓	0.0048 / 4.2120	0.0064 / 6.5220	0.0052 / 5.6910	0.0052 / 5.7520	0.0094 / 7.2310	0.0157 / 8.0780	0.0162 / 11.814	0.0158 / 10.584
HighREV-16↓	0.0053 / 4.5560	0.0083 / 7.1450	0.0059 / 6.1700	0.0060 / 6.1320	0.0114 / 7.5430	0.0158 / 9.0720	0.0173 / 11.648	0.0177 / 10.329

exposures. EBFI [1] is more efficient than ours in parameters and FLOPs, but lacks temporal modeling, leading to lower performance. Our method, while relatively heavier in parameters and FLOPs, achieves substantially better performance with comparable runtime by leveraging temporal constraints.

2.6.4 Network Details

Fig. 2.9 illustrates the detailed architectures of the encoder and decoder, which are not fully described in the main paper. The frame encoder processes the consecutive captured frames, while the event encoder extracts features from the stacked events produced by the TES module. These features are then fused via a channel-wise attention mechanism [35]. The fused representations are adaptively blended by the TIM module to produce the target feature $\hat{F}_\tau$. The decoder takes $\hat{F}_\tau$ along with the positional encoding of the target timestamp τ to reconstruct the sharp target frame $\hat{I}_\tau$. A final refinement step is applied using the fusion network adopted from [15].

2.6.5 Limitation

Although we have shown the effectiveness of our method, it has certain limitations that are worth further exploration. Specifically, the current framework assumes that frames and events share the same spatial resolution. In real-world scenarios, however, frame-based cameras and event sensors often operate at different resolutions [9]. To enhance the practicality of our approach, future work could explore strategies for fusing frame and event data with mismatched resolutions. One promising direction is to incorporate implicit neural representations for resolution-agnostic fusion [54, 55], which we leave as a potential avenue for future research.

Table 2.6: NIQE evaluations on RealBlur-DAVIS. **Bold** indicates the best score.

Dataset	NIQE↓							
	Ours	EBFI	REFID	EVDI	UEVD+TL	VIDUE	UTI	NAFNet+RIFE
RealBlur-DAVIS	26.55	28.62	29.48	38.97	37.79	36.72	40.67	41.14

Table 2.7: Computational cost on GoPro-10↓ (RandEx setting). Runtime measured on a TITAN RTX GPU. **Bold** and underlined indicate the best and second-best results, respectively.

Metric	Ours	REFID	EBFI	EVDI	UEVD+TL	VIDUE	UTI	NAFNet+RIFE
# Params (M)↓	10.58	15.76	<u>4.23</u>	0.396	107.09	62.46	49.02	70.07
FLOPs (G)↓	166.83	182.04	<u>70.08</u>	53.08	671.83	285.93	221.09	189.66
Runtime (ms)↓	<u>90.93</u>	164.60	96.23	69.20	420.40	195.40	192.30	152.40
PSNR↑	33.39	<u>31.03</u>	30.89	27.51	23.51	25.86	22.17	21.05

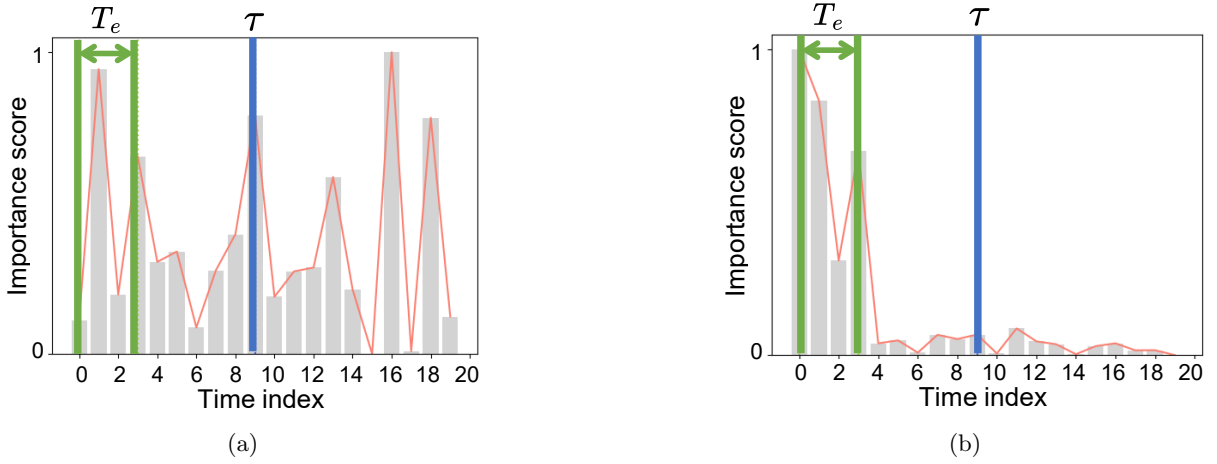


Figure 2.7: Channel-wise average importance scores $\frac{\partial \mathcal{E}_{\tau \rightarrow [t_s, t_e]}}{\partial \mathbb{E}_N}$ showing how much each temporal slice in $\mathbb{E}_N$ contributes to the sampled events. (a) Without the sampling process and (b) without the target timestamp τ . Higher scores indicate stronger selection by the TES module.

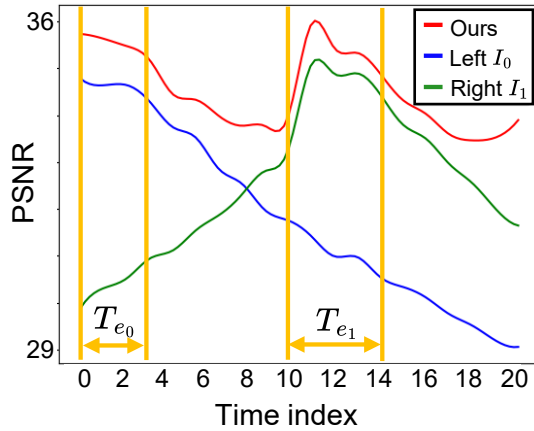


Figure 2.8: PSNR comparison over time with and without temporal constraints. Our method (red) maintains stable performance across target timestamps, while the version without temporal constraints (blue & green) shows degradation.

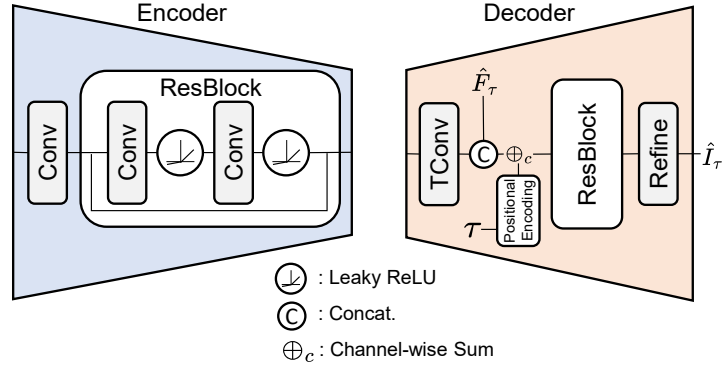


Figure 2.9: Detailed architecture of the encoder and decoder. Frame and event encoders share the same structure. The model follows a U-Net design with three encoders and two decoders, where only the final decoder includes a refinement block.

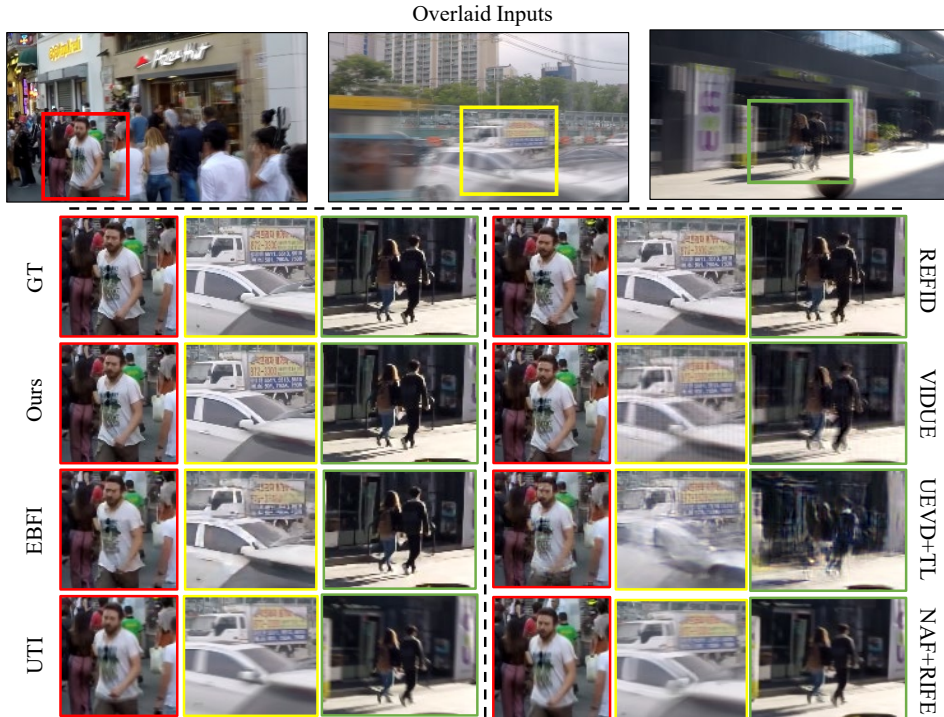


Figure 2.10: Qualitative results on GoPro (synthetic blur & events).

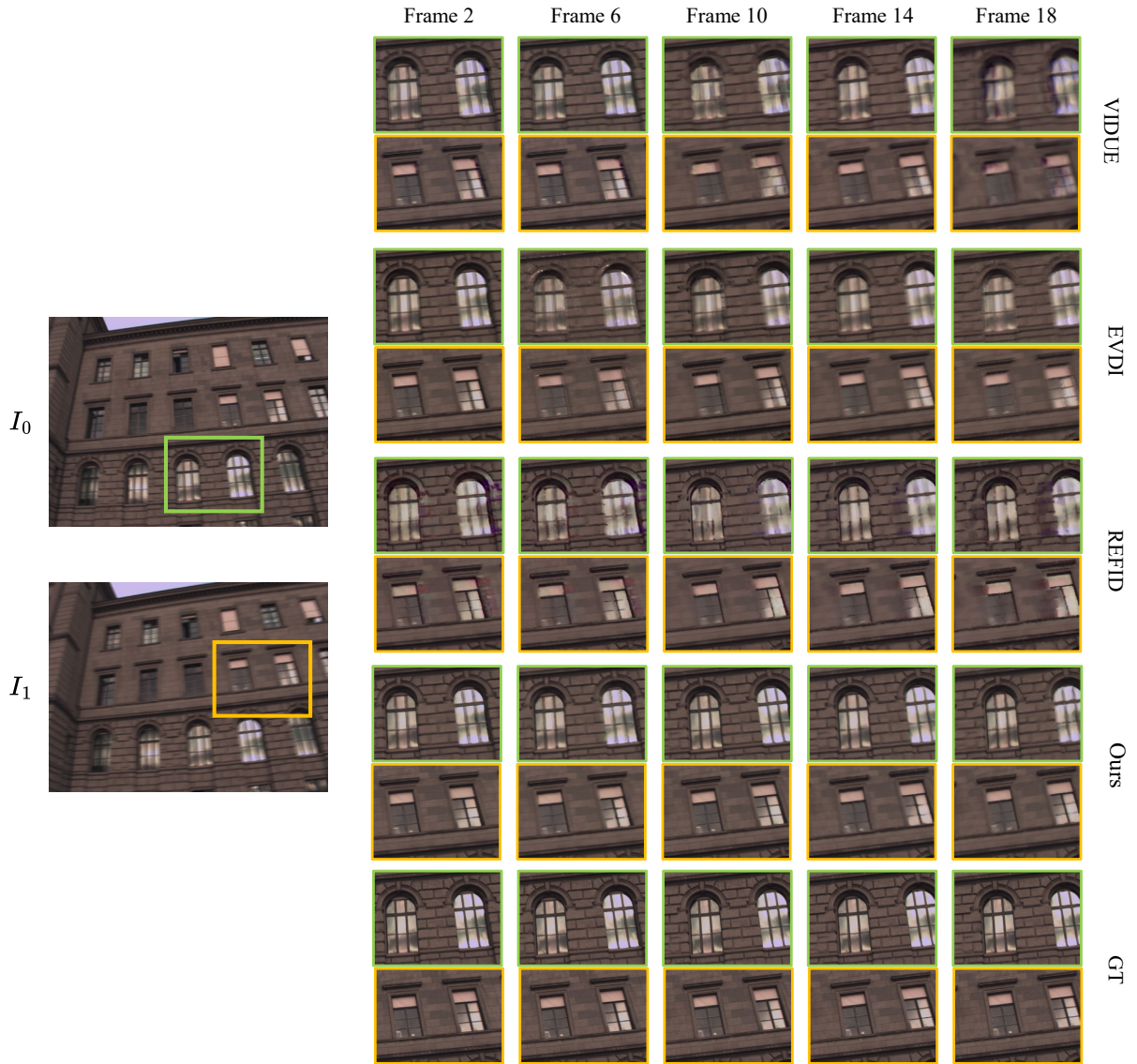


Figure 2.11: Qualitative results on HighREV (synthetic blur & real events).

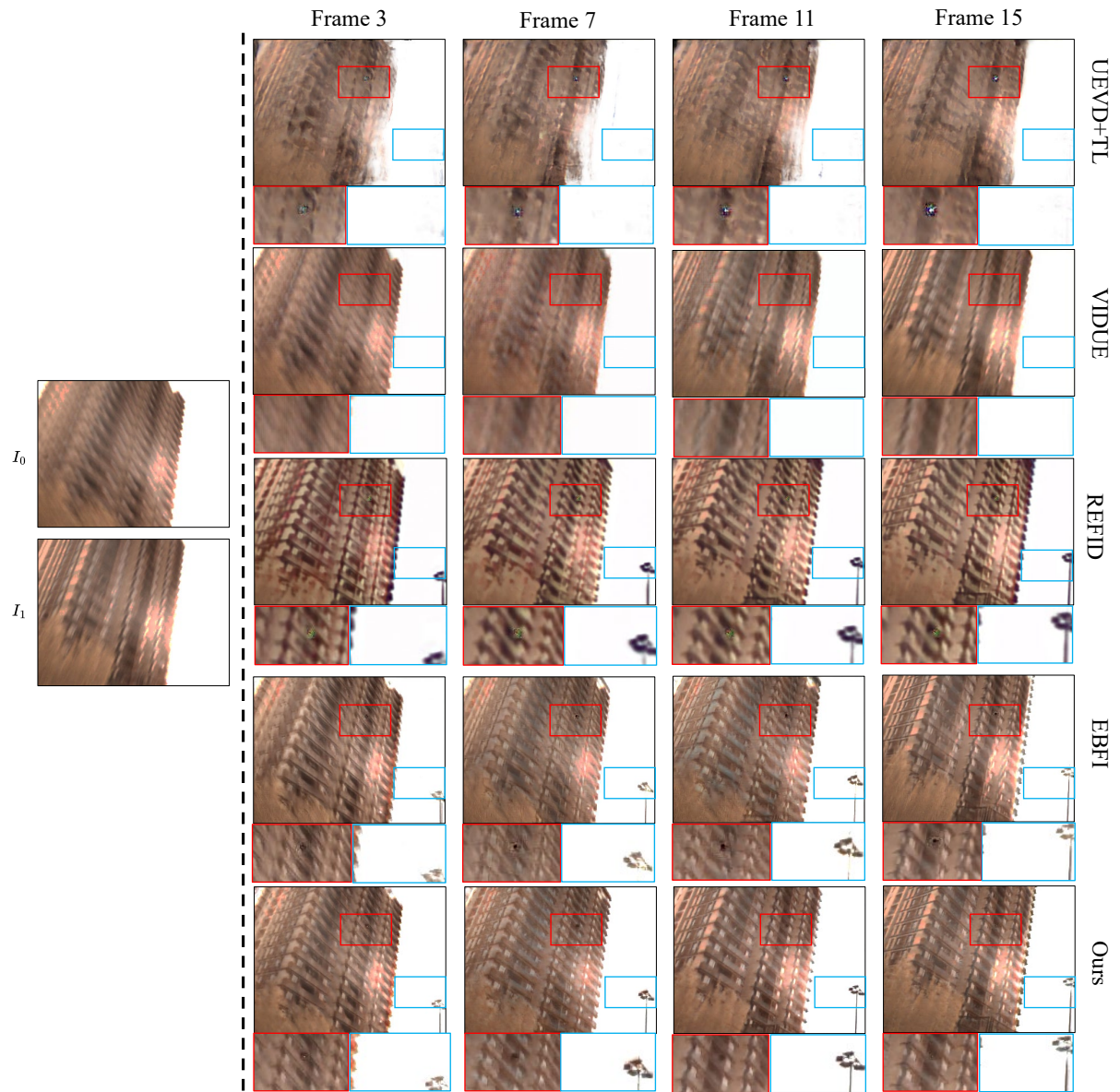


Figure 2.12: More qualitative results on RealBlur-DAVIS (real blur & events)



Figure 2.13: (continued) More qualitative results on RealBlur-DAVIS.

Chapter 3. EvBS: Event-guided Blur Synthesis for Domain-adaptive Motion Deblurring

3.1 Introduction

Motion deblurring, the task of restoring a sharp image from a blurry input caused by camera or object movement, is a fundamental challenge in computational photography. While learning-based methods have driven significant progress [8, 56, 57], their reliance on standard RGB frames inherently limits their performance in highly dynamic scenes. In conventional cameras, motion and visual content are inextricably entangled within a single exposure, causing a significant loss of motion information that makes it difficult to handle severe blur. To overcome this limitation, recent research has increasingly adopted event cameras [58]. Thanks to their microsecond-level temporal resolution and high dynamic range, event streams provide highly reliable motion cues, leading to a new paradigm of event-guided deblurring architectures [1, 35, 59].

Despite these advances, a critical challenge persists: the domain shift. Pre-trained deblurring models often exhibit significant performance degradation when applied to real-world test scenes that differ from their training distribution [60]. This discrepancy, typically arising from variations in camera environments and motion blur distributions, poses a significant risk to applications requiring reliable perception, such as autonomous driving and robotics [6]. To address this, several studies [61, 62] have introduced test-time adaptation approaches that derive supervision signals by reconstructing the input blurry image during inference. However, such indirect supervision often provides insufficient guidance for robust adaptation, restricting the model to limited update steps and yielding only marginal performance gains.

Moving a step further, recent domain-adaptive schemes [3, 4] synthesize training pairs for fine-tuning the pre-trained models by identifying pseudo-sharp patches in the test data and blurring them with their estimated motions. While promising, these methods face the following limitations: (1) They are constrained by the inherent entanglement between motion and visual content in RGB frames — meaning motion trajectories are baked into captured consecutive frames and are highly challenging to disentangle. Consequently, these methods can only re-apply a motion pattern back to its original pseudo-sharp counterpart. This inability to mix-and-match diverse motions with different contents restricts the diversity of the synthesized data, hindering robust adaptation to various real-world blurs. (2) They are designed exclusively for frame-based architectures, making them incompatible with recent event-based deblurring models [1, 5, 35, 59]. In light of the increasing adoption of events for robust multimedia perception [9], establishing adaptation frameworks for event-based architectures is highly necessary.

To overcome these challenges, we propose **EvBS**, an **Event-guided Blur Synthesis** framework that effectively generates diverse blurred patches within the target domain (See Fig. 3.1). By decoupling motion from content via the high temporal resolution of event cameras, our method not only performs *intrinsic blur* synthesis, which re-applies motion to its original content, but also enables *extrinsic blur* synthesis—transferring motion patterns from various blurred patches to given sharp contents. Specifically, EvBS comprises three main components: (1) A **Dual Source Extractor** that scans the target domain to locate sharp content sources and motion-consistent blur sources. (2) Operating directly on these detected sources, a **Blur Conditioning Module** generates blur conditions via two generators: *Intrinsic-blur Condition Generator* that extracts a sharp content source’s inherent motion, and *Extrinsic-*

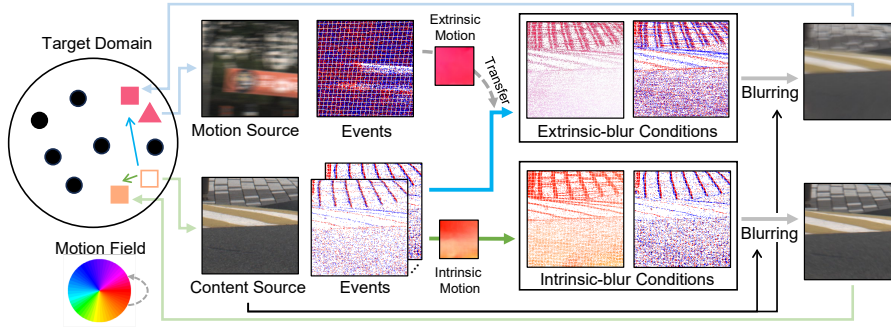


Figure 3.1: Illustration of EvBS. In the target domain representation, shapes denote visual content and colors represent distinct motion patterns. By leveraging events to explicitly decouple motion from content, EvBS enables both intrinsic blur synthesis (applying a content source’s inherent motion) and extrinsic blur synthesis (transferring motion from a separate motion source). This strategy generates a diverse set of synthesized training pairs within the target domain, facilitating enhanced model calibration.

blur Condition Generator that applies motion pattern from the blur sources onto another sharp content. (3) These conditions guide a diffusion-based model [63] to synthesize diverse training pairs for fine-tuning. By utilizing diverse motion patterns inherent in the target domain, we enable a much broader spectrum of training data synthesis. Moreover, our method is broadly applicable to both event-based and frame-based deblurring models.

Our main contributions are summarized as follows:

- We propose EvBS, a novel event-guided blur synthesis framework for domain-adaptive motion deblurring. By leveraging events to explicitly decouple motion from visual content, EvBS generates a diverse set of training pairs, breaking the inherent constraints of naturally coupled motion and content.
- We introduce a framework featuring a Dual Source Extractor and a Blur Conditioning Module to enable two complementary strategies: Intrinsic-Blur Synthesis for applying inherent motion, and Extrinsic-Blur Synthesis for transferring motion across distinct sources.
- Extensive experiments demonstrate that EvBS effectively enhances the robustness of both pre-trained image- and event-based deblurring models on unseen real-world datasets.

3.2 Related Work

3.2.1 Motion Deblurring

Early foundational works on motion deblurring have predominantly focused on CNN-based architectures [8]. These methods aim to learn a direct mapping from a blurry input to its corresponding sharp image. To enhance performance, the field has seen a variety of innovations, including the integration of attention mechanisms [64–66], coarse-to-fine strategies [56, 67], and diffusion-based approaches [57, 68]. These advancements have significantly improved the accuracy and robustness of deblurring models on established benchmarks. However, image-based approaches inherently face limitations when dealing with severe motion blur, as they struggle to leverage the rich temporal information embedded within the blur itself.

To overcome the limitations of image-based methods, recent research has shifted towards using novel sensors, most notably the event camera [58], whose high temporal resolution is particularly advantageous for motion deblurring. These approaches are generally built upon event-based image formation models [21] and integrate them with deep learning techniques to achieve remarkable efficiency and accuracy. Seminal works in this area have introduced sophisticated event representation and attention mechanisms to effectively fuse image and event features [5, 35, 59]. Furthermore, the field has progressed towards addressing more realistic and challenging constraints, such as deblurring under blind exposure times [1, 18].

Despite these advancements in both image-based and event-based methods, a common challenge persists across the field: many of these methods exhibit a significant performance drop when applied to unseen domains, particularly on real-world blurry images that differ from their training distributions. This domain gap issue motivates the need for domain adaptation techniques in motion deblurring.

3.2.2 Blur-driven Domain Adaptation

While numerous domain adaptation works have been studied for other vision tasks [69–77], their application to motion deblurring has been relatively underexplored. Among existing domain-adaptive deblurring methods, generating self-supervision signals from the test data has emerged as a dominant strategy. These methods generally fall into two categories: reconstruction-based and synthesis-based approaches.

Reconstruction-based approaches train models to reproduce the original blur during inference. Specifically, Chi *et al.* [61] and Nah *et al.* [62] proposed test-time adaptation schemes that derive supervision signals by reconstructing the input blurry images. However, this indirect supervision often provides insufficient guidance, restricting the model to a few update steps and yielding marginal performance gains. Synthesis-based approaches provide more direct supervision by generating pseudo-training pairs. Recent works [3, 4] identify pseudo-sharp patches within the target domain and synthesize training pairs by blurring them with their inherent motion patterns. While effective for fine-tuning, these methods face the following limitations. First, they are constrained by the inherent entanglement of motion and content. Because motion is naturally embedded within consecutive frames, these methods can only re-apply motion to its original pseudo-sharp counterpart, limiting the combinatorial diversity required for adapting to various blurs in the target domain. Second, relying on frame-based blurring makes them inapplicable to the well-developed event-based deblurring models [1, 5, 35, 59].

To overcome these bottlenecks, we propose leveraging the high temporal resolution of events to explicitly decouple motion from visual content. This decoupling enables both intrinsic motion synthesis (using inherent motion) and extrinsic motion transfer across different sources, generating a diverse set of training pairs beyond naturally coupled combinations, while natively supporting event-based deblurring models.

3.3 Method

Due to the inherent entanglement between motion and content, existing blur-synthesis methods are limited to generating training pairs using only the motion patterns inherent to the sharp patches themselves. To overcome this, we propose leveraging events to explicitly decouple motion from visual content. This strategy allows us to synthesize diverse training pairs by utilizing both intrinsic motion (a

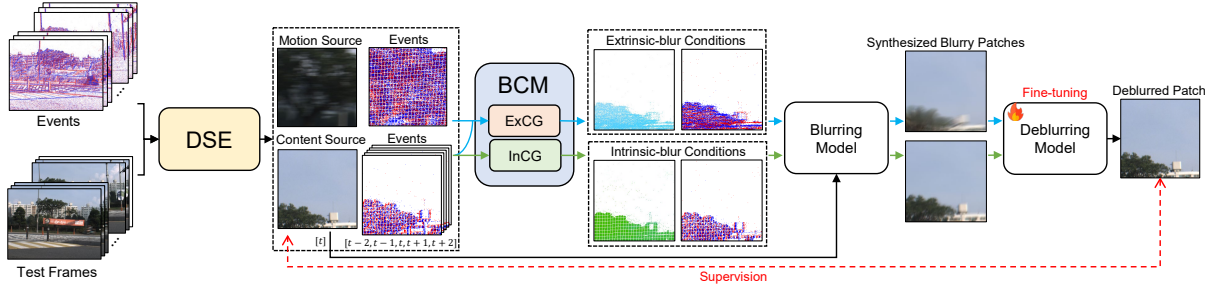


Figure 3.2: Overview of the proposed Event-guided Blur Synthesis (EvBS) framework. Given test frames and their corresponding event streams from the target domain, the Dual Source Extractor (DSE) extracts pseudo-sharp patches as content sources and blurred patches as motion sources. The Blur Conditioning Module (BCM) then generates diverse blur conditions via two pathways: the Intrinsic-blur Condition Generator (InCG) utilizes the inherent motion of the content sources, while the Extrinsic-blur Condition Generator (ExCG) transfers motions from the motion sources to different content sources. These conditions guide a blurring model to synthesize a diverse set of blurred patches. Finally, these synthesized patches are used to fine-tune a pre-trained deblurring model, with the original content sources serving as direct supervision (red dashed line).

content’s own pattern) and extrinsic motion (patterns transferred from different sources).

As illustrated in Fig. 3.2, our framework operates through two main components. First, the **Dual Source Extractor (DSE)** identifies and extracts pseudo-sharp content sources and distinct motion sources from the test data. Next, the **Blur Conditioning Module (BCM)** constructs diverse synthesis conditions through two pathways: the *Intrinsic-blur Condition Generator (InCG)* applies a content source’s own motion, whereas the *Extrinsic-blur Condition Generator (ExCG)* transfers motion patterns from the motion sources to different content sources. These conditioning signals then guide a diffusion-based reblurring model [63] to synthesize the diverse training pairs required to fine-tune the pre-trained deblurring network on the target domain. The following subsections detail each framework component.

3.3.1 Dual Source Extractor (DSE)

To synthesize diverse training pairs for target domain adaptation, we first extract suitable content sources (pseudo-sharp patches) and motion sources (patches with clear motion patterns) from the given test frames. As shown in Fig. 3.3, our DSE achieves this through a blur-aware patch detection approach. **Blur score estimation.** We leverage the observation that regions with high event activity typically correspond to areas with significant motion. From the input event stream, we generate an event count map $\mathcal{C}_E$ that accumulates the number of events at each pixel location. To complement this, we employ a Blur Magnitude Estimator (BME) [4], which predicts the spatial distribution of blur intensity and produces a blur magnitude map $\mathcal{M}$. Unlike the original BME that uses only RGB images, we enhance its estimation capability by retraining it with both the input image and the corresponding $\mathcal{C}_E$. This multi-modal integration allows the model to better capture motion-intensive regions. We then compute a comprehensive Blur Score Map $\mathcal{B}$ by combining these two complementary cues:

$$\mathcal{B} = \alpha \cdot \mathcal{C}_E + \beta \cdot \mathcal{M}, \quad (3.1)$$

where α and β are weighting parameters. Higher values in $\mathcal{B}$ indicate blurrier regions (motion source candidates), while lower values suggest sharper regions (content source candidates).

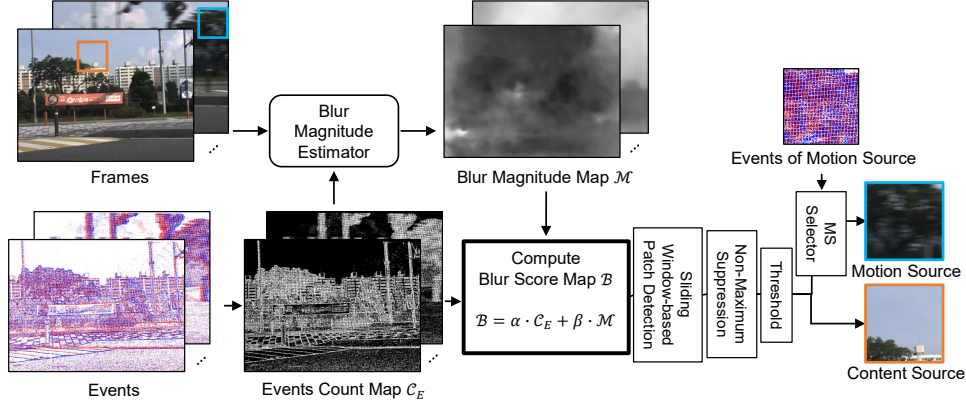


Figure 3.3: Illustration of the Dual Source Extractor. The blur score map $\mathcal{B}$ is computed from the event count map $\mathcal{C}_E$ and the blur magnitude map $\mathcal{M}$. Based on this map, sliding-window detection identifies content sources (low scores) and motion source candidates (high scores). Subsequent non-maximum suppression (NMS) and thresholding steps are utilized to eliminate overlapping patches and reduce redundancy. Motion sources are further filtered by the motion source selector based on motion consistency.

Patch detection and refinement. Using the blur score map $\mathcal{B}$, we extract content and motion source candidates from regions where $\mathcal{B} < \tau$ and $\mathcal{B} > 1 - \tau$, respectively. The threshold τ is adaptively adjusted [4] so that each source type constitutes approximately $r\%$ of the total frames (ablated in Sec. 3.4.3). Following sliding-window detection and non-maximum suppression to remove overlapping patches, we further refine the motion source candidates. To ensure reliable motion transfer, we introduce a Motion Source (MS) Selector that evaluates the directional coherence of candidate patches. It explicitly filters out patches with chaotic or conflicting trajectories, retaining only those with highly coherent motions as the final sources.

3.3.2 Blur Conditioning Module (BCM)

Once content and motion sources are extracted, the Blur Conditioning Module (BCM) generates multi-modal motion conditions to guide the subsequent blur synthesis. For the synthesis process, we build upon ID-Blau [63]. Let $\mathcal{D}_\psi$ denote this conditional diffusion model parameterized by ψ , which was originally designed to generate a blurred image B given a sharp content image C and a flow map $\mathcal{F} = [\tilde{u}; \tilde{v}; m] \in \mathbb{R}^{3 \times H \times W}$ (encoding unit orientation $(\tilde{u}, \tilde{v})$ and magnitude m):

$$B = \mathcal{D}_\psi(\mathcal{F}, C). \quad (3.2)$$

While the flow map $\mathcal{F}$ effectively captures macro-level motion trajectories, it lacks the temporal granularity required to model local micro-motion dynamics. To overcome this, our BCM leverages event data as an auxiliary conditioning signal. Specifically, we introduce Flow-Enhanced Deviation Accumulation (FEDA) to supplement the macro motion of $\mathcal{F}$ with the fine-grained, high-frequency dynamics embedded in events

The deviation accumulation representation [5] effectively captures pure motion contours by aggregating the temporal variations of event streams. Building upon this, FEDA advances the representation by weighting the events with the magnitude derived from the corresponding motion vector. This approach preserves the overall motion trajectory while emphasizing regions with high motion intensity (detailed

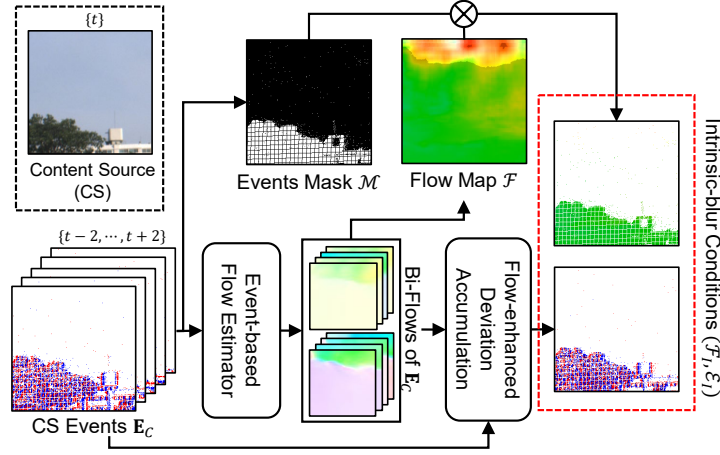


Figure 3.4: Intrinsic-blur Condition Generator. Given events $\mathbf{E}_C$ from the temporal window $[t-2, t+2]$ around a content source at frame t , we first estimate bidirectional optical flows. These flows are aggregated into a flow map and then multiplied by an event mask $\mathcal{M}$ to produce the final intrinsic flow map $\mathcal{F}_I$. Concurrently, $\mathbf{E}_C$ is processed via Flow-Enhanced Deviation Accumulation (FEDA) to generate the intrinsic event representation $\mathcal{E}_I$. Together, the pair $(\mathcal{F}_I, \mathcal{E}_I)$ serves as the intrinsic-blur conditions.

in the Supplementary Material). Given this event-guided FEDA representation $\mathcal{E} \in \mathbb{R}^{H \times W}$, we train a modified $\mathcal{D}_\psi$ from scratch to synthesize the final blurred image B :

$$B = \mathcal{D}_\psi(\mathcal{F}, \mathcal{E}, C). \quad (3.3)$$

This refined, event-guided conditioning improves the domain-specific blur synthesis, ultimately boosting the efficacy of the fine-tuning process (see Sec. 3.4.3).

To formulate the synthesis conditions $(\mathcal{F}$ and $\mathcal{E})$, the BCM comprises two complementary generators: the Intrinsic-blur Condition Generator (InCG) and the Extrinsic-blur Condition Generator (ExCG). Both generators operate on the shared principle of explicitly decoupling motion from visual content by extracting motion cues exclusively from event streams. Building upon this decoupled cues, InCG constructs conditioning signals using the content source’s own motion patterns. In contrast, ExCG formulates conditions by transferring the motion patterns of an external motion source to a distinct content source.

Intrinsic-blur Condition Generator (InCG). InCG generates conditions for blurring each content source using its own motion pattern, as illustrated in Fig. 3.4. Given a content source from the t -th frame, we utilize the corresponding events $\mathbf{E}_C$ from the temporal window $[t-2, t+2]$ to capture the local motion dynamics. These events are fed into an event-based optical flow estimator [78] to extract forward flows $\{\mathbf{F}_{t-2 \rightarrow t-1}, \dots, \mathbf{F}_{t+1 \rightarrow t+2}\}$ and backward flows $\{\mathbf{F}_{t-1 \rightarrow t-2}, \dots, \mathbf{F}_{t+2 \rightarrow t+1}\}$.

We then compute a flow map by aggregating these bidirectional flows [63]:

$$\mathcal{F} = \sum_{n=t-2}^{t+1} \frac{\mathbf{F}_{n \rightarrow n+1} - \mathbf{F}_{n+1 \rightarrow n}}{2},$$

where $\mathcal{F} = [u; v] \in \mathbb{R}^{2 \times H \times W}$ represents the horizontal and vertical motion components. For conditioning, we reformulate this into a 3-channel tensor $\mathcal{F} = [\tilde{u}; \tilde{v}; \tilde{m}] \in \mathbb{R}^{3 \times H \times W}$, where $\tilde{u}$ and $\tilde{v}$ encode the normalized motion orientation, and $\tilde{m}$ encodes the normalized magnitude:

$$\tilde{u} = \frac{u}{\sqrt{u^2 + v^2}}, \quad \tilde{v} = \frac{v}{\sqrt{u^2 + v^2}}, \quad \tilde{m} = \frac{\sqrt{u^2 + v^2}}{M}, \quad (3.4)$$

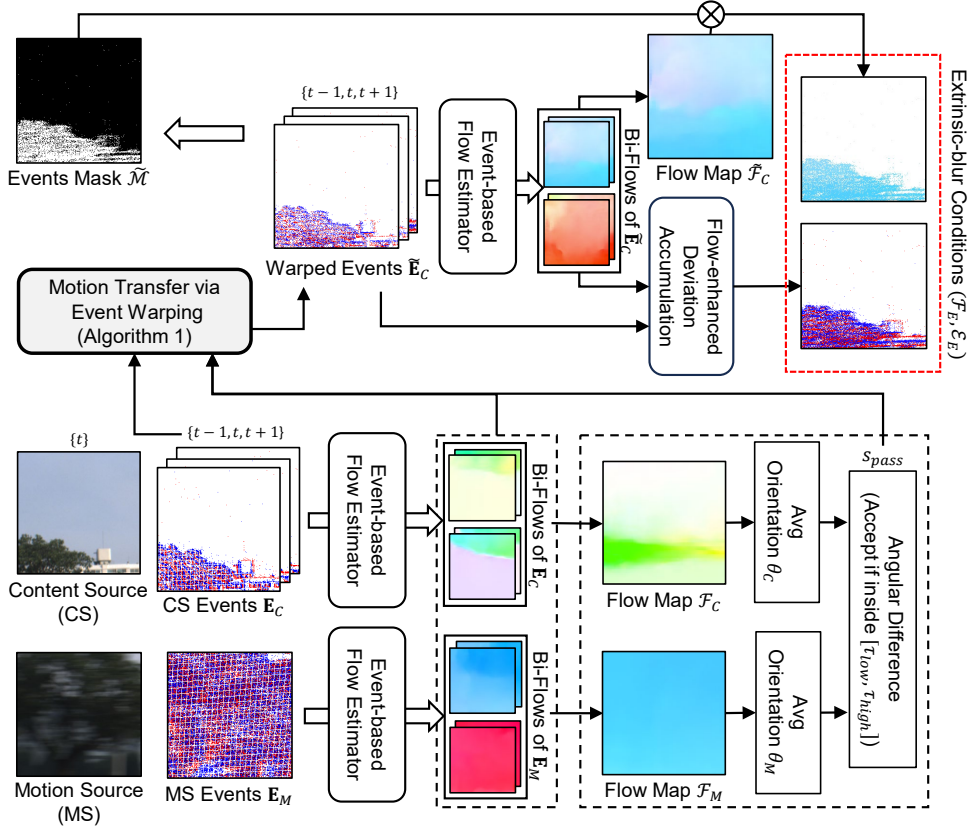


Figure 3.5: Extrinsic-blur Condition Generator. Flow maps $\mathcal{F}_C$ and $\mathcal{F}_M$ are first extracted from content events $\mathbf{E}_C$ and motion events $\mathbf{E}_M$. If their angular difference falls within $[\tau_{low}, \tau_{high}]$, the motion pattern from $\mathbf{E}_M$ is transferred to $\mathbf{E}_C$ via Algorithm 1 to produce the warped events $\tilde{\mathbf{E}}_C$. Finally, $\tilde{\mathbf{E}}_C$ is processed via FEDA to yield the extrinsic flow map $\mathcal{F}_E$ and event representation $\mathcal{E}_E$. The resulting condition $(\mathcal{F}_E, \mathcal{E}_E)$ effectively integrates the content’s spatial structure with external motion patterns.

where M denotes the maximum magnitude within the current map. To focus exclusively on motion-active regions, we create a binary mask $\mathcal{M} \in \{0, 1\}^{H \times W}$, where $\mathcal{M}(x, y) = 1$ if at least one event occurred at (x, y) , and 0 otherwise. We apply this mask to obtain the final intrinsic flow map $\mathcal{F}_I = \mathcal{F} \odot \mathcal{M}$, where $\odot$ denotes element-wise multiplication. Concurrently, we process $\mathbf{E}_C$ via the FEDA algorithm to generate the intrinsic event representation $\mathcal{E}_I \in \mathbb{R}^{H \times W}$.

Consequently, the pair $(\mathcal{F}_I, \mathcal{E}_I)$ serves as the intrinsic-blur condition, effectively guiding the blurring model to synthesize blur that is governed by the content source’s inherent motion.

Extrinsic-blur Condition Generator (ExCG). As illustrated in Fig. 3.5, the ExCG formulates conditioning signals by transferring extrinsic motion patterns from distinct motion sources to a target content source. Given a content source from the t -th frame, we extract corresponding events from the narrower temporal window $[t-1, t+1]$ to minimize misalignment during motion transfer. Using an event-based flow estimator [78], we compute the bidirectional flows and aggregate them into a content flow map $\mathcal{F}_C = \sum_{n=t-1}^t (\mathbf{F}_{n \rightarrow n+1} - \mathbf{F}_{n+1 \rightarrow n})/2$. Concurrently, for a given motion source spanning the temporal interval $[t_s, t_e]$, we divide its continuous event stream into three consecutive temporal segments. We then extract the bidirectional flows between these sequential segments and aggregate them to construct a motion flow map $\mathcal{F}_M = \sum_{i=1}^2 (\mathbf{F}_{i \rightarrow i+1} - \mathbf{F}_{i+1 \rightarrow i})/2$, where i denotes the segment indices. Here, both flow maps $\mathcal{F}_C, \mathcal{F}_M \in \mathbb{R}^{2 \times H \times W}$ encode the pixel-wise motion vectors $[u; v]$.

Algorithm 1 Motion Transfer via Event Warping

```

1: Input:
   Content source events  $\mathbf{E}_C$ ,
   Bi-Flows of  $\mathbf{E}_C$ :  $\{\mathbf{F}_{t-1 \rightarrow t}, \mathbf{F}_{t \rightarrow t+1}\}, \{\mathbf{F}_{t \rightarrow t-1}, \mathbf{F}_{t+1 \rightarrow t}\}$ .
   Bi-Flows of  $\mathbf{E}_M$ :  $\{\mathbf{F}_{i-1 \rightarrow i}, \mathbf{F}_{i \rightarrow i+1}\}, \{\mathbf{F}_{i \rightarrow i-1}, \mathbf{F}_{i+1 \rightarrow i}\}$ .
   Similarity check result  $s_{\text{pass}} \in \{\text{true}, \text{false}\}$ 
2: Output: Warped events  $\tilde{\mathbf{E}}_C$ 
3: if  $s_{\text{pass}} = \text{false}$  then
4:   return  $\emptyset$ 
5: end if
6: Get median time  $t_{\text{med}}$  from  $\mathbf{E}_C$  at  $t$ 
7: for each direction  $k \in \{1, -1\}$  do
8:   // Get directional flows
9:    $\mathcal{F}_C^k \leftarrow (\mathbf{F}_{t \rightarrow t+k} - \mathbf{F}_{t+k \rightarrow t})/2$ 
10:   $\mathcal{F}_M^k \leftarrow (\mathbf{F}_{i \rightarrow i+k} - \mathbf{F}_{i+k \rightarrow i})/2$ 
11:  // Extract local & global motion (Eq. 3.6)
12:   $\mathcal{F}_C^k \rightarrow \mathbf{V}_C^k, \mathcal{F}_M^k \rightarrow \mathbf{v}_M^k$ 
13:  // Compute offset field
14:   $\mathbf{T}_k \leftarrow \mathbf{v}_M^k - \mathbf{V}_C^k$ 
15:  // Split events  $\mathbf{E}_C$ 
16:   $\mathbf{E}_k \leftarrow \mathbf{E}_C$  in  $[t_{\text{med}}, t+1]$  if  $k = 1$ , else  $[t-1, t_{\text{med}}]$ 
17:  // Set bounds
18:   $[t_{\min}, t_{\max}] \leftarrow [t_{\text{med}}, t+1]$  if  $k = 1$ , else  $[t-1, t_{\text{med}}]$ 
19:  for each event  $(x, y, \tau) \in \mathbf{E}_k$  do
20:     $w(\tau) \leftarrow \frac{1-k}{2} + k \cdot \frac{\tau - t_{\min}}{t_{\max} - t_{\min}}$ 
21:     $(x', y') \leftarrow (x, y) + w(\tau) \cdot \mathbf{T}_k(x, y)$ 
22:    Add  $(x', y', \tau)$  to  $\tilde{\mathbf{E}}_C$ 
23:  end for
24: end for
25: return Warped events  $\tilde{\mathbf{E}}_C$ 

```

To identify motion sources that offer sufficient diversity while remaining transferable, we evaluate their orientation similarity with the target content source. First, we compute the average orientation vectors $\boldsymbol{\theta}_C, \boldsymbol{\theta}_M \in \mathbb{R}^2$ from $\mathcal{F}_C$ and $\mathcal{F}_M$. For a given flow map, the average components are calculated as:

$$\bar{u} = \frac{1}{N} \sum u, \quad \bar{v} = \frac{1}{N} \sum v, \quad \boldsymbol{\theta} = \frac{[\bar{u}; \bar{v}]}{\|[\bar{u}; \bar{v}]\|}, \quad (3.5)$$

where the summation is computed over all valid pixels in the flow map, and N denotes the total number of such pixels. We then compute the angular difference $\phi = \arccos(\boldsymbol{\theta}_C \cdot \boldsymbol{\theta}_M)$ between the content orientation and motion orientation. We proceed with the motion transfer only if this angle falls within a specific range, $\pi_{\text{low}} < \phi < \pi_{\text{high}}$ (ablated in Sec. 3.4.3). This filtering excludes redundant motion sources that may limit diversity and highly disparate ones that cause unrealistic synthesized blur.

Once a motion source passes the similarity check, we transfer its motion pattern to the content source via event warping, as detailed in Algorithm 1. The core idea is to compute an offset field $\mathbf{T}_k(x, y)$ that represents the difference between the target motion source’s global motion and the original content

source’s local motion.

Specifically, for each temporal direction ($k \in \{1, -1\}$), we first compute the bidirectionally aggregated flow maps $\mathcal{F}_C^k = [u_c; v_c]$ and $\mathcal{F}_M^k = [u_m; v_m] \in \mathbb{R}^{2 \times H \times W}$. For the content source, we directly use the flow map $\mathcal{F}_C^k$ as the local motion map $\mathbf{V}_C^k$. To extract the global motion vector $\mathbf{v}_M^k$ from $\mathcal{F}_M^k$, we spatially average its components across N valid pixels: $\bar{u}_m = \frac{1}{N} \sum u_m$, $\bar{v}_m = \frac{1}{N} \sum v_m$, and $\bar{m} = \frac{1}{N} \sum m$, where $m = \sqrt{u_m^2 + v_m^2}$. Then, the local motion map $\mathbf{V}_C^k \in \mathbb{R}^{2 \times H \times W}$ and the global motion vector $\mathbf{v}_M^k \in \mathbb{R}^2$ are formally defined as:

$$\mathbf{V}_C^k = [u_c; v_c], \quad \mathbf{v}_M^k = \frac{[\bar{u}_m; \bar{v}_m]}{\|[\bar{u}_m; \bar{v}_m]\|} \cdot \bar{m}. \quad (3.6)$$

We then compute the offset field $\mathbf{T}_k \in \mathbb{R}^{2 \times H \times W}$ by subtracting the local motion map from the spatially broadcasted global motion vector:

$$\mathbf{T}_k = \mathbf{v}_M^k - \mathbf{V}_C^k. \quad (3.7)$$

Finally, each event in $\mathbf{E}_C$ at timestamp τ is warped using $\mathbf{T}_k$, scaled by a temporal weight $w(\tau)$ proportional to its distance from the $\mathbf{E}_C$ ’s median time t_{med} . This operation effectively neutralizes the original inherent motion pattern in $\mathbf{E}_C$ while injecting the target extrinsic pattern, resulting in the warped events $\tilde{\mathbf{E}}_C$.

Using the warped events $\tilde{\mathbf{E}}_C$, we extract new bidirectional flows $\tilde{\mathbf{F}}$ and a corresponding event mask $\tilde{\mathcal{M}}$ to compute the extrinsic flow map $\mathcal{F}_E = (\sum_{n=t-1}^t (\tilde{\mathbf{F}}_{n \rightarrow n+1} - \tilde{\mathbf{F}}_{n+1 \rightarrow n}) / 2) \odot \tilde{\mathcal{M}}$. Concurrently, $\tilde{\mathbf{E}}_C$ is processed via FEDA to yield the extrinsic event representation $\mathcal{E}_E$. Ultimately, the pair $(\mathcal{F}_E, \mathcal{E}_E)$ serves as the extrinsic-blur condition, guiding the model to synthesize blur with the transferred motion pattern.

3.3.3 Blur Synthesis

The generated intrinsic and extrinsic conditions are fed into the modified ID-Blau $\mathcal{D}_\psi$ to synthesize the final blurred images B_t^I and B_t^E for a given content source C_t :

$$B_t^I = \mathcal{D}_\psi(\mathcal{F}_I, \mathcal{E}_I, C_t), \quad B_t^E = \mathcal{D}_\psi(\mathcal{F}_E, \mathcal{E}_E, C_t). \quad (3.8)$$

This process yields diverse blur-sharp training pairs tailored for target domain adaptation. Specifically, when fine-tuning an event-based deblurring model, it takes the synthesized blurred patch (either B_t^I or B_t^E) alongside its corresponding events (either $\mathbf{E}_C$ or $\tilde{\mathbf{E}}_C$) as inputs to restore the original pseudo-sharp content C_t .

Extrinsic pair selection. When multiple motion sources satisfy the orientation criterion, we prioritize the top- n pairs that best preserve motion consistency. To quantify this, we compute a motion consistency score $\text{MCS} = (\tilde{\boldsymbol{\theta}}_C \cdot \boldsymbol{\theta}_M + 1) / 2$ between the average orientation vectors $\tilde{\boldsymbol{\theta}}_C, \boldsymbol{\theta}_M \in \mathbb{R}^2$ derived from $\tilde{\mathcal{F}}_C$ and $\mathcal{F}_M$ via Eq. 3.5. We then select the top- n motion sources with the highest MCS, where n is determined through ablation studies (Sec. 3.4.3).

3.4 Experiments

3.4.1 Experimental Setups

Datasets. We retrain the Blur Magnitude Estimator (BME) [4] and ID-Blau [63] on the GoPro [2] training set with synthetic events [46]. For evaluation, we employ benchmarks covering both real and synthesized event-based deblurring datasets. Real-world performance is assessed on EVRB [79] and

Table 3.1: Quantitative results across deblurring benchmarks. We compare baseline performance with results obtained after adaptation via our method (w/ Ours). * indicates event-based models.

Models	EVRB			REVD			HighREV-3			HighREV-7			HighREV-11		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
MAENet*	25.46	0.791	0.428	25.41	0.813	0.421	36.35	0.939	0.303	31.79	0.859	0.366	27.24	0.771	0.447
w/ Ours	27.93	0.851	0.386	28.64	0.858	0.372	39.66	0.965	0.245	33.91	0.929	0.309	29.92	0.832	0.391
EFNet*	25.77	0.797	0.419	26.39	0.827	0.394	37.34	0.942	0.293	32.35	0.873	0.359	27.77	0.791	0.432
w/ Ours	28.17	0.859	0.375	28.96	0.865	0.363	40.23	0.977	0.212	34.51	0.942	0.305	30.01	0.844	0.379
EBFI*	25.31	0.785	0.435	26.14	0.829	0.402	38.22	0.943	0.299	31.93	0.860	0.377	27.33	0.778	0.442
w/ Ours	27.51	0.845	0.389	28.46	0.862	0.357	40.67	0.970	0.242	34.65	0.941	0.309	29.85	0.835	0.397
NAFNet	24.13	0.746	0.473	24.74	0.794	0.452	36.16	0.917	0.322	29.76	0.828	0.415	25.57	0.738	0.486
w/ Ours	26.24	0.817	0.419	26.55	0.838	0.399	38.53	0.961	0.268	31.95	0.861	0.358	27.63	0.784	0.426
MIMO-UNet	23.97	0.732	0.487	24.34	0.789	0.459	36.69	0.922	0.326	29.79	0.835	0.406	25.01	0.727	0.491
w/ Ours	26.01	0.807	0.423	26.18	0.825	0.401	38.11	0.959	0.277	32.03	0.871	0.349	27.23	0.772	0.431

Table 3.2: Comparison with recent domain-adaptive deblurring methods. Best results in **bold**.

Models	EVRB		REVD		HighREV-3		HighREV-7		HighREV-11	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
NAFNet	24.13	.746	24.74	.794	36.16	.917	29.76	.828	25.57	.738
+ MetaDeblur	24.22	.753	24.81	.797	36.54	.929	29.97	.832	25.60	.739
+ DADeblur	25.23	.781	25.87	.819	37.71	.946	31.28	.849	27.19	.767
+ Ours	26.24	.817	26.55	.838	38.53	.961	31.95	.861	27.63	.784
MIMO-UNet	23.97	.732	24.34	.789	36.69	.922	29.79	.835	25.01	.727
+ MetaDeblur	24.05	.736	24.43	.795	36.87	.935	29.92	.842	25.02	.727
+ DADeblur	25.18	.774	25.63	.811	37.77	.943	31.09	.856	26.87	.755
+ Ours	26.01	.807	26.18	.825	38.11	.959	32.03	.871	27.23	.772

REVD [80], which provide real motion blur and events from hybrid cameras. To overcome the limited availability of real-world event-based deblurring datasets, we utilize HighREV [16]. Using the realistic blur synthesis pipeline in [81], we generate three variants — HighREV-3, HighREV-7, and HighREV-11 — by aggregating 3, 7, and 11 consecutive frames, respectively, to represent a comprehensive spectrum of blur intensities.

Implementation Details. We employ E-RAFT [78] for event-based flow estimation. For BME [4] and ID-Blau [63], we follow their original training settings while modifying their input spaces: BME is adapted to accept both RGB images and event count maps $\mathcal{C}_E$, and ID-Blau is updated to take the flow map $\mathcal{F}$, FEDA representation $\mathcal{E}$, and content image C as conditions.

We evaluate our framework by fine-tuning five recent GoPro-pretrained deblurring models, including event-based (EFNet [35], MAENet [5], EBFI [1]) and frame-only (NAFNet [37], MIMO-UNet [56]) architectures. All models are fine-tuned for 10 epochs on our synthesized pairs using the Charbonnier loss [45] and Adam optimizer [82], with the learning rate cosine-annealed [48] from 10^{-5} to 10^{-7} . We report PSNR, SSIM [49], and LPIPS [50]. Key hyperparameters (ablated in Sec. 3.4.3) include the source selection ratio $r = 0.15$, the number of extrinsic pairs $n = 2$, and orientation angle thresholds $[\tau_{\text{low}}, \tau_{\text{high}}] = [15^\circ, 30^\circ]$. Empirical settings include: a patch size of 256×256 with a stride of 64 for sliding window-based patch detection, initial patch threshold $\tau = 0.05$, NMS IoU threshold of 0.3, and blur score weights $\alpha = \beta = 0.5$ (Eq. 3.1).

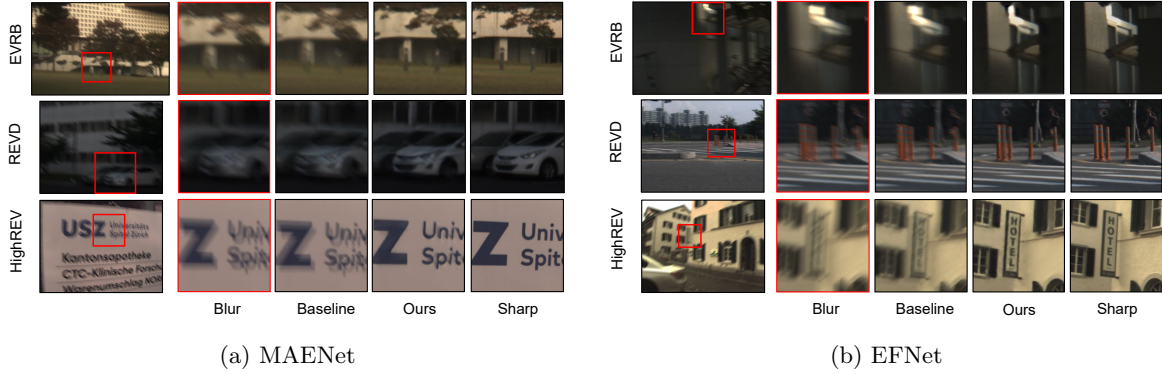


Figure 3.6: Qualitative results across diverse models and datasets. We compare the blurry input (blur), pre-trained baseline (baseline), our adapted model (ours), and ground truth (sharp), with red boxes denoting zoomed-in regions. Our approach consistently restores finer details across both event-based deblurring models. Additional visual results for the other models (EBFI, NAFNet, MIMO-UNet) are shown in the supplementary material.

3.4.2 Experimental Results

Quantitative Results. Table 3.1 presents the quantitative evaluation on various deblurring benchmarks. Our method consistently boosts performance across diverse architectures, including event-based (MAENet, EFNet, EBFI) and frame-only models (NAFNet, MIMO-UNet). Specifically, we achieve substantial average gains of 2.24 dB on EVRB, 2.35 dB on REVD, and up to 2.63 dB on HighREV datasets. These results validate the generalizability of our approach, demonstrating its capability to effectively adapt pre-trained models to diverse target domains.

Table 3.2 compares our method with recent domain-adaptive blur synthesis approaches [3, 4]. As these existing methods cannot process event streams, we conduct the comparison using frame-only backbones (NAFNet and MIMO-UNet). Our method consistently yields superior results across diverse models and datasets. On NAFNet, we achieve notable gains over the best baseline: 1.01 dB on EVRB, 0.68 dB on REVD, and up to 0.82 dB on the HighREV datasets. MIMO-UNet shows similar improvements of 0.83 dB (EVRB), 0.55 dB (REVD), and up to 0.94 dB on HighREV. These findings validate that leveraging both intrinsic and extrinsic motion cues effectively enhances domain-adaptive deblurring.

Qualitative Results. Fig. 3.6 provides qualitative comparisons on the EVRB, REVD, and HighREV datasets using two event-based models (MAENet, EFNet). Due to space constraints, additional visual results for the other architectures (EBFI, NAFNet, MIMO-UNet) are shown in the supplementary material. Compared to the pre-trained baselines, our adapted models consistently reconstruct sharper images with superior detail preservation. As highlighted in the magnified crops, our method effectively restores challenging high-frequency structures—such as legible text and sharp object edges—that remain blurred in the baseline outputs. These improvements validate the efficacy of our approach, demonstrating that synthesizing diverse training pairs via intrinsic and extrinsic motion transfer leads to robust target domain adaptation. Furthermore, Fig. 3.7 compares our framework against recent domain-adaptive methods (MetaDeblur [3], DADeblur [4]) using NAFNet [37] as the baseline. The results demonstrate that our method consistently achieves better restoration of fine details on unseen domains compared to the recent domain-adaptive approaches.

Fig. 3.8 demonstrates the qualitative results of our dual-pathway blur synthesis framework. For a

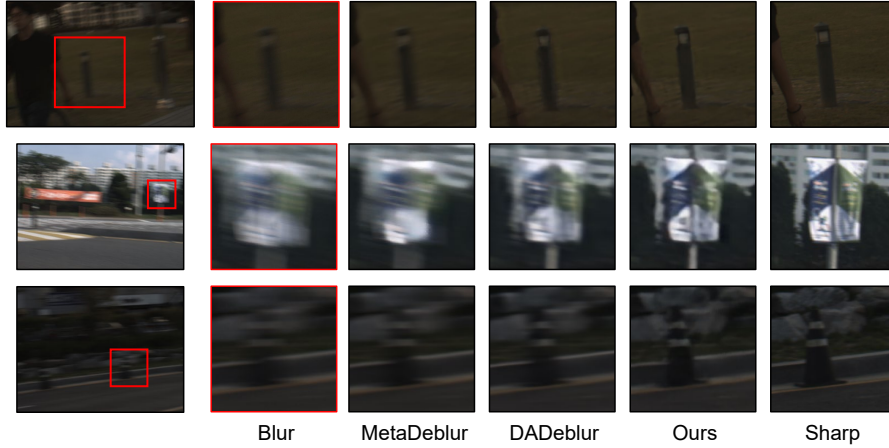


Figure 3.7: Visual comparison of deblurring results against recent domain-adaptive methods [3, 4]. Our method consistently yields sharper and more visually pleasing results across diverse scenes.

given content source, our approach comprehensively generates training pairs by leveraging both intrinsic and extrinsic motion sources. Specifically, while it utilizes the content’s inherent flow map $\mathcal{F}_C$ for intrinsic blur synthesis (green box), it expands the diversity by generating extrinsic blur utilizing the transferred flow map $\tilde{\mathcal{F}}_C$ (blue box, derived from a separate motion source $\mathcal{F}_M$). By unlocking this dual generation capability via motion decoupling, we can synthesize a wide spectrum of distinct blurred patches from a single sharp content source, expanding the target domain’s data diversity. Notably, the transferred flow $\tilde{\mathcal{F}}_C$ structurally aligns with the motion source pattern $\mathcal{F}_M$ rather than the content source $\mathcal{F}_C$, validating our effective transfer mechanism.

3.4.3 Ablation Studies

We conduct ablation studies on the REVD [80] using MAENet [5].

Effects of r , n , and $[\tau_{low}, \tau_{high}]$. Fig. 3.9 analyzes the impact of key hyperparameters. First, the source selection ratio peaks at $r = 0.15$; lower values yield insufficient training pairs, while higher ratios introduce label noise by selecting patches with residual blur. Second, $n = 2$ extrinsic pairs yields the best performance. Increasing n incorporates lower-ranked (low MCS) candidates with weaker alignment to the target motion pattern, thereby compromising adaptation efficacy. Lastly, evaluating the orientation threshold reveals that the $[\tau_{low}, \tau_{high}] = [15^\circ, 30^\circ]$ range provides an effective balance between directional consistency and motion diversity. A narrower range may restrict diversity by filtering out valid motion sources, while a broader range introduces disparate motions that induce unrealistic transfer artifacts.

Effectiveness of Motion Components. Table 3.3 summarizes the contributions of our Blur Conditioning Module (BCM). Compared to the pre-trained baseline (25.41 dB), fine-tuning with solely InCG (intrinsic) or ExCG (extrinsic) pairs improves PSNR to 28.03 dB and 28.14 dB, respectively. Combining both yields the best performance of 28.64 dB, demonstrating the synergistic effect of utilizing diverse motion priors in the target domain.

To validate the ExCG design, we test a random offset baseline by randomly sampling $\mathbf{T}_k$ (Algorithm 1). This drops the PSNR to 24.58 dB, even underperforming the baseline. This degradation highlights that our motion-aware transfer is crucial for synthesizing pairs aligned with the target domain distribution.

Effectiveness of Event-guided Blurring. We validate our FEDA representation against two condi-

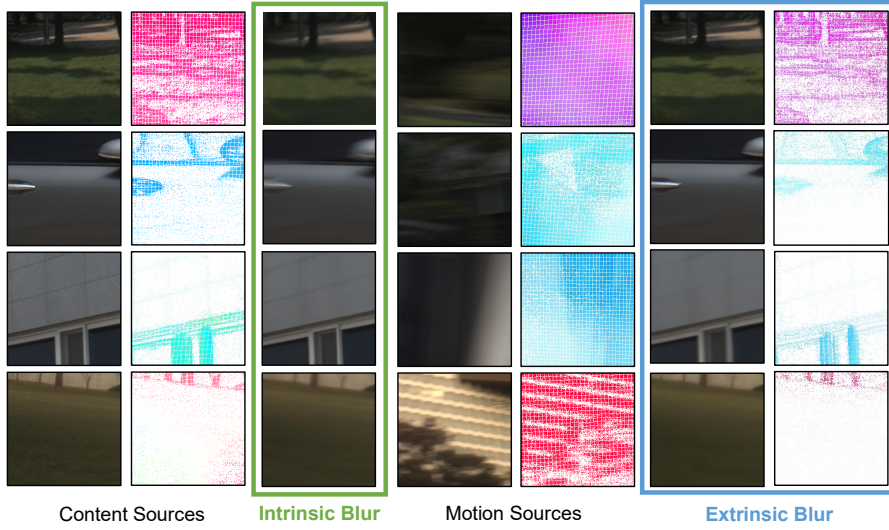


Figure 3.8: Qualitative examples of intrinsic and extrinsic blur synthesis. For each row, we display the content source and its flow map $\mathcal{F}_C$ (left), followed by the **intrinsic synthesized blur**. To demonstrate motion transfer, we show a separate motion source and its flow map $\mathcal{F}_M$ (middle), and the resulting **extrinsic synthesized blur** with the transferred flow map $\tilde{\mathcal{F}}_C$ (right). All flow maps are masked by event masks to explicitly highlight motion-intensive regions.

tioning baselines: using only flow maps (w/o Events, as in ID-Blau [63]), and with original deviation accumulation (w/ DA [5]) on GoPro test set. As shown in Fig. 3.10, flow-only conditioning fails to accurately capture motion trajectories, whereas the DA captures motion trajectories but lacks capturing motion intensity. Conversely, our FEDA accurately models both trajectory and intensity, yielding the best blur-synthesizing quality (36.82 dB in PSNR across GoPro test set) that closely matches the ground truth.

Lastly, when these synthesized pairs are used for domain-adaptive motion deburring, our FEDA-guided model achieves the highest performance (28.64 dB), compared to the adapted models trained with DA (28.53 dB) and flow-only (28.16 dB) conditionings. This confirms that FEDA provides richer motion cues, producing more effective training pairs for fine-tuning.

Table 3.3: Effectiveness of motion components. Combining both intrinsic and extrinsic motion yields the highest PSNR. Conversely, applying random motion offsets severely degrades performance, validating the necessity of our motion transfer algorithm.

Intrinsic	Extrinsic	Random	PSNR (dB)
			25.41
✓			28.03
	✓		28.14
		✓	24.58
✓	✓		28.64

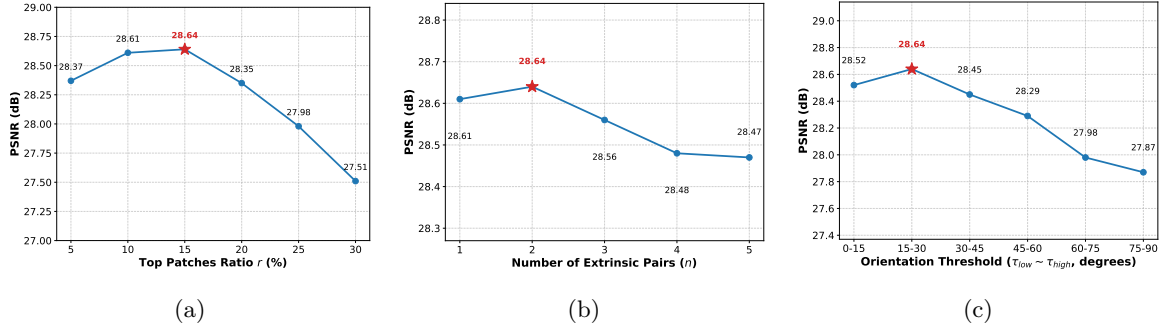


Figure 3.9: Ablation study on key hyperparameters. We analyze the impact of (a) the top patches ratio r , (b) the number of extrinsic pairs n , and (c) the orientation threshold ranges $[\tau_{low}, \tau_{high}]$. The best values are marked with red stars.

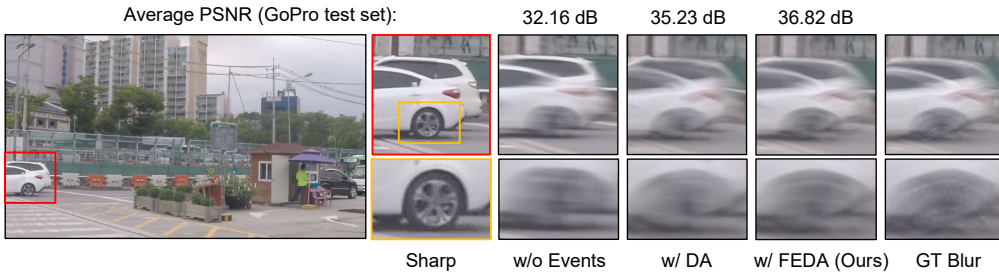


Figure 3.10: Visual comparison of blur synthesis under different conditioning. While the baseline using only flow maps (w/o Events) only captures macro-motions, and DA [5] struggles in motion-intensive areas, our FEDA successfully synthesizes blur patterns with high fidelity. The average PSNR scores on the GoPro test set (top) quantitatively show its superior blur synthesis quality.

3.5 Conclusion

In this paper, we propose EvBS, a novel event-guided diverse blur synthesis framework, to address the domain gap in motion deblurring. By leveraging events to decouple motion information from visual content, our framework synthesizes diverse training pairs through intrinsic and extrinsic motion synthesis strategies. These strategies produce richer training data that break the inherent constraints of naturally coupled motion and content. Comprehensive experimental results demonstrate that EvBS effectively enhances adaptation performance, significantly improving the robustness of state-of-the-art deblurring models on unseen domains.

3.6 Appendix

3.6.1 Flow-Enhanced Deviation Accumulation (FEDA)

Deviation Accumulation (DA) [5] effectively encodes motion trajectory but accumulates events uniformly, disregarding motion intensity. To generate richer event-based conditioning signals, we propose FEDA, which enhances DA with quadratic flow [83] magnitude information to better capture fine-grained motion dynamics.

Algorithm 2 details the FEDA process. Unlike standard deviation accumulation, FEDA weights each event by its corresponding motion magnitude. Specifically, for each event at (x, y) , we estimate

Algorithm 2 Flow-Enhanced Deviation Accumulation

0: **Input:** Taking a pixel at (x, y) as a specific example, N Events stream $\{\mathbf{E}_0, \mathbf{E}_1, \dots, \mathbf{E}_{N-1}\}$, Bidirectional flows $\{\mathbf{F}_{0 \rightarrow 1}, \mathbf{F}_{1 \rightarrow 2}, \dots, \mathbf{F}_{N-1 \rightarrow N}\}$, $\{\mathbf{F}_{0 \rightarrow -1}, \mathbf{F}_{1 \rightarrow 0}, \dots, \mathbf{F}_{N-1 \rightarrow N-2}\}$.
0: **Output:** FEDA event representation V .
1: $R = 0, D = 0, C = 0$
2: **for** $i = 0$ to $N - 1$ **do**
3: Counting the number of events C_i at (x, y)
4: **for** $k = 0$ to $C_i - 1$ **do**
5: Event $E_i = (x, y, t_k, p_k)$
6: Normalize timestamps: $dt_k = \frac{t_k - t_0}{t_{C_i-1} - t_0}$
7: Compute $\mathbf{v}_0 = \frac{\mathbf{F}_{i \rightarrow i+1} - \mathbf{F}_{i \rightarrow i-1}}{2}$
8: Compute $\mathbf{a} = \mathbf{F}_{i \rightarrow i+1} + \mathbf{F}_{i \rightarrow i-1}$
9: Quadratic flow: $\mathbf{Q}_{i \rightarrow t_k} = \mathbf{v}_0 \cdot dt_k + 0.5 \cdot \mathbf{a} \cdot dt_k^2$
10: Extract flow magnitude $|M_{t_k}|$ from $\mathbf{Q}_{i \rightarrow t_k}$
11: $D = p_k \cdot |M_{t_k}|$
12: $R = R + D$
13: **end for**
14: $C = C + C_i$
15: **end for**
16: Compute final FEDA representation: $V = R/C$

the motion vector using quadratic flow, which explicitly models acceleration. The event polarity is then scaled by this computed motion magnitude ($|M_{t_k}|$) and accumulated. Finally, the sum is normalized by the total event count to yield a robust event representation V . This magnitude-weighted accumulation emphasizes motion-intensive regions more effectively, producing discriminative event representations.

3.6.2 Motion Source Selector (MS Selector)

To ensure directional coherence in motion transfer, we introduce a Motion Source (MS) Selector (Fig. 3.11) that evaluates the consistency of motion patterns present in the given motion source. Specifically, we partition the event stream into two consecutive temporal segments and estimate the optical flow $\mathbf{F} = [u; v] \in \mathbb{R}^{2 \times H \times W}$ using event-based flow estimator [78].

Using the horizontal and vertical components u, v , we compute the average orientation $\boldsymbol{\theta}_{\text{avg}} \in \mathbb{R}^2$ following the formulation in Eq. 3.5:

$$\bar{u} = \frac{1}{N} \sum u, \quad \bar{v} = \frac{1}{N} \sum v, \quad \boldsymbol{\theta}_{\text{avg}} = \frac{[\bar{u}; \bar{v}]}{\|[\bar{u}; \bar{v}]\|}, \quad (3.9)$$

where N is the number of pixels. Concurrently, we identify the dominant orientation $\boldsymbol{\theta}_{\text{max}} = [u; v]_{\mathbf{x}^*} \in \mathbb{R}^2$ at the location $\mathbf{x}^*$ with the maximum flow magnitude ($\|\mathbf{F}_{\mathbf{x}}\|_2$). Finally, we calculate the cosine similarity $S = \boldsymbol{\theta}_{\text{avg}} \cdot \boldsymbol{\theta}_{\text{max}}$. We accept candidates only if $S > \tau$ (with $\tau = 0.9$), ensuring that the selected source exhibits a coherent trajectory suitable for transfer.

3.6.3 More Qualitative Results

We provide additional qualitative comparisons on EBF1 [1], NAFNet [37], and MIMO-UNet [56] in Fig. 3.12. While the baseline model struggles with residual blur and artifacts due to domain shift,

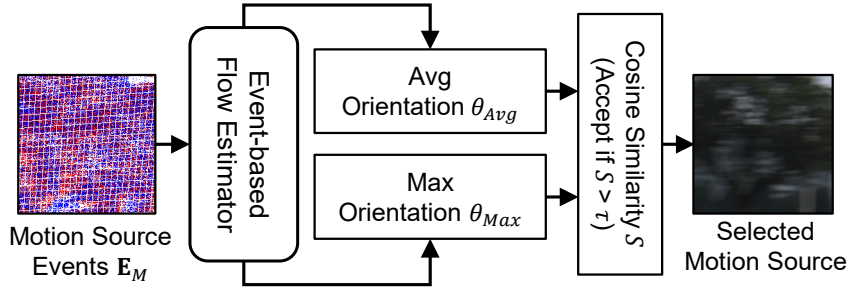


Figure 3.11: Illustration of the Motion Source Selector

Table 3.4: Computational cost analysis (adapting on EVRB with MAENet, using RTX 2080 Ti).

Component	Time (s)
Dual Source Extractor	270.3
Blur Conditioning Module	
Intrinsic Condition Generator	61.5
Extrinsic Condition Generator	428.8
Blurring	288.7
Fine-tuning	264.3

our adapted version consistently produces sharper images with better-preserved fine details. This visual evidence further validates the robustness of our event-guided domain adaptation framework across diverse real-world datasets.

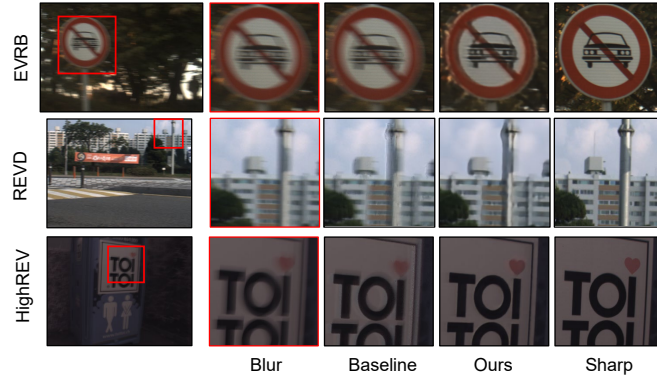
Furthermore, in Fig. 3.13, we present additional visual examples of our blur synthesis process. These results illustrate our framework’s ability to consistently generate diverse blurred patches across various scenes. As shown, our method successfully synthesizes both intrinsic blur, utilizing the content’s inherent motion, and extrinsic blur, driven by motion patterns transferred from separate motion sources. This further confirms that our motion decoupling mechanism effectively expands the diversity of training pairs, which ultimately drives the enhanced adaptation performance.

3.6.4 Computational Cost

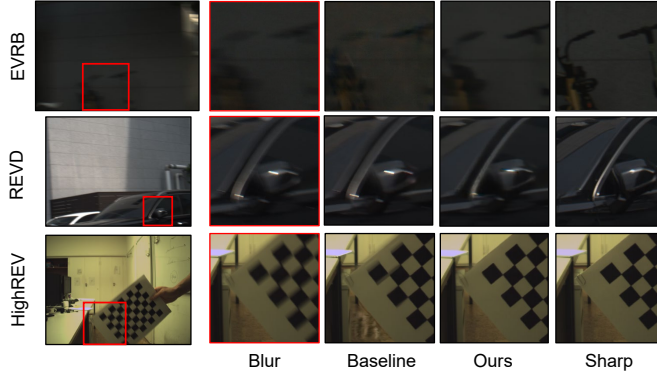
Table 3.4 details the computational cost of our framework on the EVRB dataset (894 images) using MAENet [5] with a single NVIDIA RTX 2080 Ti. The Dual Source Extractor takes 270.3s, while the Motion Conditioning Module requires 490.3s in total. Notably, the Extrinsic Condition Generator (ExCG) dominates the cost (428.8s) due to the computational complexity of motion pattern matching and transfer. The blurring and fine-tuning stages take 288.7s and 264.3s, respectively. Finally, it is worth noting that our adaptation process is conducted offline; thus, it does not affect the run-time efficiency of the deblurring model during inference.

3.6.5 Limitation

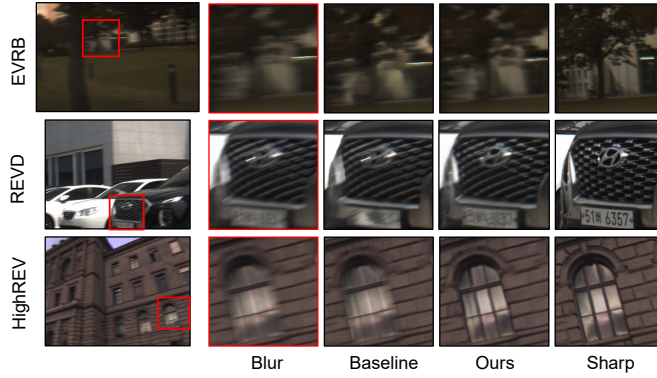
One consideration in our framework is the choice of the source selection ratio r . Increasing r raises the likelihood of including content sources with residual blur. Transferring extrinsic motion to such



(a) EBF1



(b) NAFNet



(c) MIMO-UNet

Figure 3.12: Additional qualitative results on EBF1, NAFNet, MIMO-UNet. We compare the blurry input (blur), pre-trained baseline (baseline), our adapted model (ours), and ground truth (sharp), with red boxes denoting zoomed-in regions. Our approach consistently restores finer details across both event-based (a) and frame-only (b, c) models.

inputs may generate unnatural blur patterns, potentially acting as noisy labels that can lower adaptation performance (Fig. 3.9a). Therefore, when applying our method to unseen datasets, we recommend setting r to a conservative range (e.g., 10–15%) to ensure the sharpness of content sources.

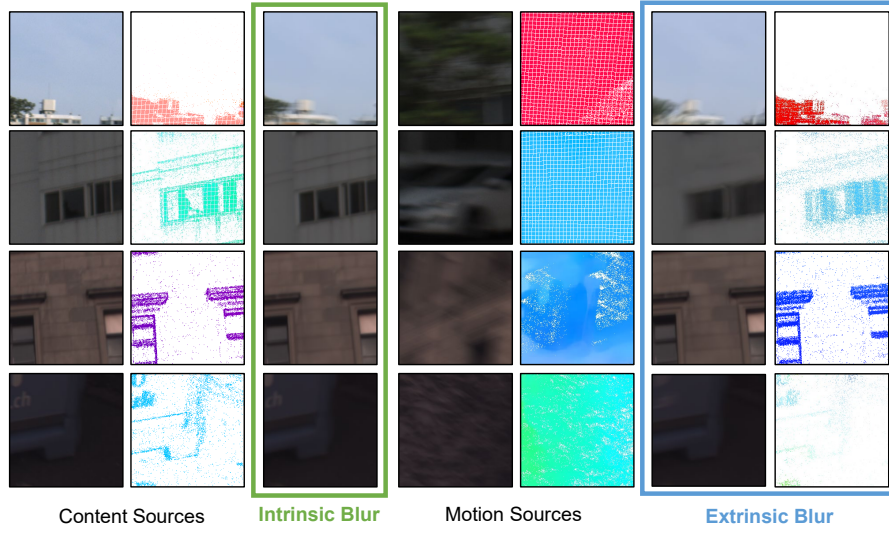


Figure 3.13: Additional qualitative examples of dual blur synthesis (intrinsic and extrinsic). Similar to Fig. 3.8 in the main manuscript, each row displays a content source (left), followed by the intrinsic synthesized blur (green box) which utilizes inherent motion. The middle columns show the motion source, and lastly, the resulting extrinsic synthesized blur on the right (blue box).

Chapter 4. EvZVD: Event-guided Zero-shot Video Deblurring Using Pre-trained Image Deblurring Diffusion Model

4.1 Introduction

Diffusion models have recently revolutionized the field of image restoration, demonstrating its capabilities in generating high-fidelity, realistic textures. Driven by strong priors learned from large datasets, pre-trained Image Diffusion Models (IDMs) excel at recovering sharp details from severely blurred images [84, 85]. Along this line, extending these powerful generative capabilities to video deblurring has become a objective in computational photography [86].

However, training dedicated Video Diffusion Models (VDMs) requires massive video datasets and substantial computational resources [86–88]. As an alternative, applying pre-trained IDMs to video frames in a zero-shot manner offers an efficient, training-free solution [87]. Yet, because IDMs process each frame independently, their inherent stochasticity often causes identical blurred regions to be rendered into different sharp textures across frames. This independent generation leads to temporal inconsistency, degrading the overall perceptual quality. While recent work [89] addressed temporal flickering by injecting cross-frame attention guided by optical flow, estimating accurate flow directly from blurred frames is challenging, limiting its effectiveness for deblurring tasks.

To address this limitation, we leverage the unique physical properties of event cameras. Unlike standard frame-based cameras, event cameras asynchronously capture pixel-level brightness changes, enabling relatively reliable motion estimation directly from event streams [78] even under motion blur where frame-based estimation frequently fails. Our key insight is that the spatial density of these events can serve as a continuous reliability score for this motion estimation.

Building on this insight, we propose a simple yet effective plug-and-play, zero-shot video deblurring framework that enhances the temporal consistency of IDMs without any additional training. Our approach begins with a globally applied initial latent warping. Subsequently, we introduce Event Reliability-Aware Spatially Adaptive Guidance, which operates at three complementary levels of the diffusion process: Cross-frame Attention propagates temporally consistent visual features, Noise Consistency Guidance (NCG) aligns the per-step stochastic noise trajectory, and Prediction Consistency Guidance (PCG) anchors the clean image prediction via gradient-based optimization. Notably, all three mechanisms are spatially modulated by the reliability score. In motion-reliable dynamic regions, we enforce consistency using warped representations from the previous frame. In static or flow-unreliable regions, we directly apply the unwrapped previous representations. This spatially adaptive modulation effectively enhancing temporal consistency of pre-trained IDMs.

Our main contributions are summarized as follows:

- We propose a simple yet effective, training-free video deblurring framework that leverages event cameras to extend pre-trained IDMs to the video domain, without compromising their robust restoration capabilities.
- We introduce Event Reliability-Aware Spatially Adaptive Guidance, which leverages event density as a continuous reliability score to adaptively modulate warp-based temporal transfer in event-dense regions and flow-free consistency constraints in event-sparse regions.

- Experiments demonstrate that our method effectively improves temporal consistency without any fine-tuning.

4.2 Related Work

Diffusion Models for Image Restoration. Diffusion models [90] have demonstrated remarkable success in image restoration. Early methods [91, 92] tackled deblurring as a linear inverse problem assuming a known blur kernel. Because acquiring accurate motion trajectories is difficult in practice, recent works [84, 93] have advanced to real-world blind deblurring, directly recovering sharp details without predefined kernels. Despite their success in single-image restoration, these IDMs process frames independently. Consequently, their direct application to video sequences inevitably causes temporal inconsistency due to the stochastic nature of diffusion, bottlenecking their use in video applications.

Video Diffusion Models for Video Restoration. To address temporal inconsistency, recent works [94, 95] train dedicated Video Diffusion Models (VDMs) by integrating temporal modules into base diffusion architectures. While explicitly modeling temporal dynamics effectively mitigates flickering, training these VDMs demands prohibitive computational resources, massive video datasets, and extensive optimization time. This extreme computational bottleneck severely restricts their practical adoption, underscoring the critical need for an efficient, training-free alternative that guarantees temporal consistency while fully exploiting pre-trained diffusion priors.

Zero-Shot Video Adaptation with Image Diffusion Models. To bypass the heavy training burden of VDMs, recent studies explore zero-shot adaptation, leveraging cross-frame attention to extend IDMs to temporally coherent video generation [87, 88] and restoration [89]. Notably, ZVRD [89] enforces temporal consistency via optical flow-guided attention without additional training. However, they rely on optical flow estimated directly from blurry frames, making its temporal guidance inherently unreliable under real-world blur. In contrast, we leverage event cameras — which capture per-pixel brightness changes independent of exposure — to obtain blur-robust flow, enabling spatially-adaptive temporal guidance via event density as a continuous reliability weight.

4.3 Method

We present a zero-shot video deblurring framework that extends pre-trained Image Diffusion Models (IDMs) to video without fine-tuning. Leveraging event cameras for blur-robust flow estimation, we propose Event Reliability-Aware Spatially Adaptive Guidance, which uses event density as a continuous spatial reliability weight to modulate three complementary temporal constraints across each denoising step: cross-frame attention for structural coherence, noise consistency guidance for step-level alignment, and prediction consistency guidance for coarse-level correction. By spatially adapting all constraints via event reliability rather than imposing them uniformly, our framework effectively suppresses temporal flickering without compromising the generative capacity of the underlying IDM.

4.3.1 Preliminary

Diffusion Models for Blind Deblurring. Recent IDM-based methods address real-world blind deblurring by operating in a VAE-compressed latent space [84, 93]. Given a clean image x_0 , the encoder maps it to $z_0 = \mathcal{E}(x_0)$, and the forward process progressively corrupts it with Gaussian noise to yield

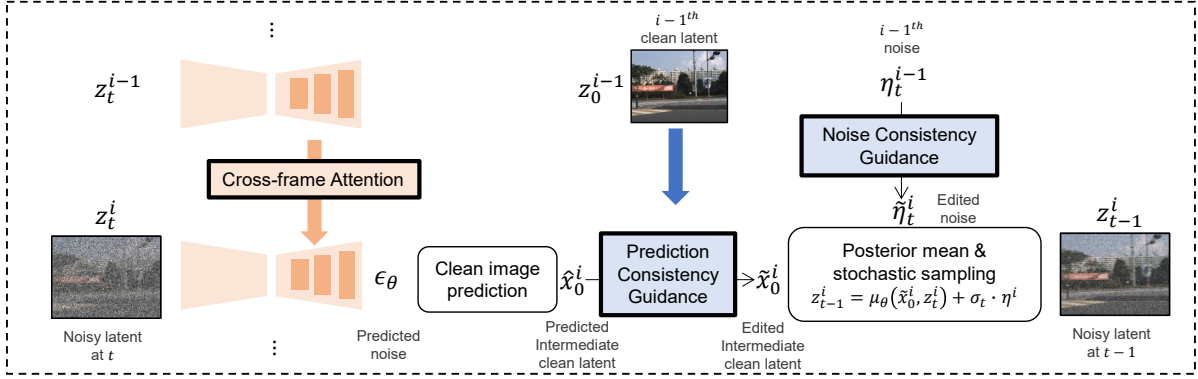


Figure 4.1: Overview of our framework. At each denoising step t , the noisy latent z_t^i is processed by the diffusion UNet with Cross-frame Attention, which injects warped Key/Value features from the previous frame z_{t-1}^{i-1} to propagate temporal context. The predicted noise ϵ_θ yields an intermediate clean latent $\hat{x}_0^i$, which is refined by Prediction Consistency Guidance toward the previous frame’s fully denoised latent z_0^{i-1} . Separately, Noise Consistency Guidance replaces the stochastic noise with a reliability-weighted blend from the previous frame’s noise η_t^{i-1} . The edited latent $\tilde{x}_0^i$ and noise $\tilde{\eta}_t^i$ are combined via posterior mean sampling to produce z_{t-1}^i .

z_t . During the reverse process, a denoising network ϵ_θ recovers the sharp latent, heavily conditioned by specialized modules that inject estimated blur priors—such as spatially adaptive kernels or kernel-aware control networks—directly into the IDM. To handle highly complex, spatially varying motion blur, some methods further reformulate the forward process as progressive blurring and enforce cross-time consistency:

$$\min_{\theta} \mathbb{E}_{t, z_0} \|f_\theta(z_t, t) - z_0\|_2^2, \quad (4.1)$$

enabling robust blind deblurring in as few as a single step. Despite their strong per-frame performance, these models process each frame independently, inevitably causing temporal flickering when applied to videos. Our method addresses this without any architectural modification. By formulating event density as a continuous spatial weight, we adaptively modulate complementary temporal constraints in both the attention mechanisms and the latent space.

4.3.2 Event Reliability-Aware Spatially Adaptive Guidance

Overview. Figure 4.1 illustrates an overview of our framework, which extends pre-trained IDMs for video deblurring without retraining. We leverage event cameras to obtain blur-robust optical flow, from which we derive event density $r \in [0, 1]$ as a continuous spatial reliability score and an occlusion mask M_{occ} . We first warp the initial noisy latent z_T to establish global temporal correlation. Building on this, we apply three complementary guidance mechanisms, each spatially modulated by r and M_{occ} throughout the reverse diffusion process: Cross-frame Attention blends attention features with warped counterparts from the previous frame for structural coherence; Noise Consistency Guidance (NCG) replaces per-step stochastic noise with a reliability-weighted blend of the warped previous noise for step-level alignment; and Prediction Consistency Guidance (PCG) applies gradient-based correction toward the warped previous clean prediction $\hat{x}_0$ to suppress low-frequency temporal drift.

Event-guided Spatial Reliability Estimation. We design two complementary spatial signals to modulate temporal guidance during denoising: a continuous reliability weight and a binary occlusion

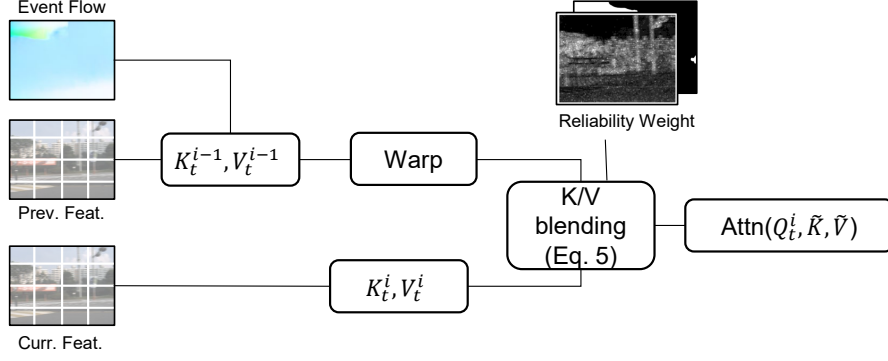


Figure 4.2: Cross-frame Attention. The Key/Value features of the previous frame $\mathbf{K}_t^{i-1}, \mathbf{V}_t^{i-1}$ are warped using event-based optical flow and blended with the current frame’s features $\mathbf{K}_t^i, \mathbf{V}_t^i$ weighted by the spatial reliability weight r . The resulting blended Keys and Values $\tilde{\mathbf{K}}, \tilde{\mathbf{V}}$ are used in the attention computation alongside the current frame’s Query $\mathbf{Q}_t^i$, propagating temporally consistent structural context into the denoising process.

mask. Given raw events between frames $t - 1$ and t , we first accumulate per-pixel event counts. Since events fire primarily at moving intensity boundaries, we dilate the count map via max-pooling to propagate the motion prior to object interiors. The dilated count map $C(p)$ is then normalized into a density weight:

$$D(p) = \frac{C(p)}{\max_p C(p)} \in [0, 1]. \quad (4.2)$$

We exclude occluded pixels from temporal guidance since they lack reliable historical correspondences. Using optical flow $\mathbf{F}$ from an event-based estimator [78], we detect occlusions via forward-backward cycle-consistency [96] at pixel p :

$$M_{\text{occ}}(p) = \mathbf{1} [|\mathbf{F}_{t \rightarrow t-1}(p) + \mathbf{F}_{t-1 \rightarrow t}(\mathbf{F}_{t \rightarrow t-1}(p))| > \delta(p)], \quad (4.3)$$

where $\delta(p) = 0.01 (|\mathbf{F}_{t-1 \rightarrow t}(p)| + |\mathbf{F}_{t \rightarrow t-1}(p)|) + \tau$ is an adaptive threshold.

Finally, we combine both signals into a unified spatial reliability weight:

$$r(p) = D(p) \cdot (1 - M_{\text{occ}}(p)) \in [0, 1]. \quad (4.4)$$

Note that all masks are computed at the image resolution and bilinearly interpolated to the latent spatial resolution $(H/8 \times W/8)$ prior to application.

Cross-frame Attention. The pipeline of cross-frame attention is shown in Fig. 4.2. Given i^{th} frame, we inject warped Key/Value (K/V) features from the previous frame into the current frame’s attention layers throughout the denoising process to propagate temporal information while preserving the model’s deblurring capacity. At each denoising step $t \in \{0, \dots, T-1\}$, the keys (and similarly values $\mathbf{V}$) are updated as:

$$\tilde{\mathbf{K}}_t^i = (1 - m) \cdot \mathbf{K}_t^i + m \cdot \mathcal{W}(\mathbf{K}_t^{i-1}, \mathbf{F}_{i-1 \rightarrow i}) \quad (4.5)$$

where $\mathcal{W}(\cdot, \mathbf{F})$ denotes the flow-based warping operator, and $\mathbf{K}_t^{i-1}$ denotes the key features of the previous frame at the corresponding denoising step, cached during its inference. The blending weight $m = r \cdot s(t)$ combines the spatial reliability weight and a step scalar $s(t)$. We adopt a linearly decreasing schedule $s(t) = 1 - t/(T-1)$, assigning stronger temporal injection early on and decaying it as denoising progresses [97]. Soft blending avoids full feature replacement, preserving the blur-conditioned features $(\mathbf{K}_t^i, \mathbf{V}_t^i)$ to maintain the model’s inherent deblurring capacity.

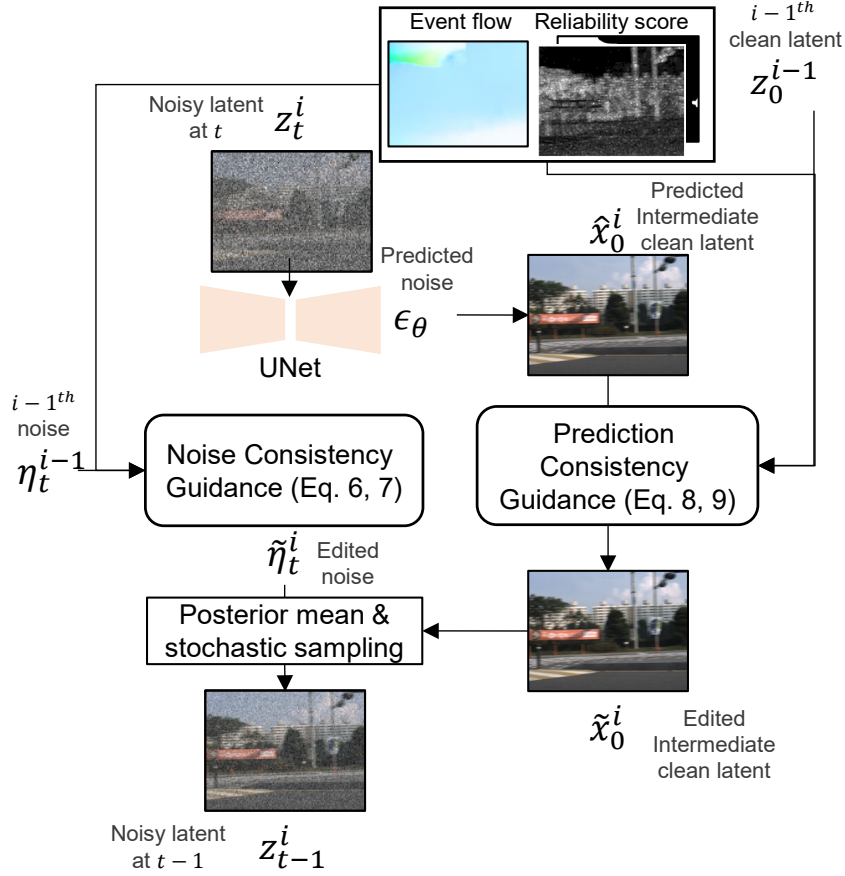


Figure 4.3: Noise and Prediction Consistency Guidance. At each denoising step, the UNet takes noisy latent z_t^i and outputs predicted noise ϵ_θ , from which an intermediate clean latent $\hat{x}_0^i$ is estimated. Prediction Consistency Guidance refines $\hat{x}_0^i$ toward the previous frame’s clean latent z_0^{i-1} , spatially weighted by the reliability score. Noise Consistency Guidance replaces the stochastic noise with a reliability-weighted blend from the previous frame’s noise η_t^{i-1} . The edited $\tilde{x}_0^i$ and $\tilde{\eta}_t^i$ are combined via posterior mean sampling to produce z_{t-1}^i .

Noise and Prediction Consistency Guidance. The pipeline of cross-frame attention is shown in Fig. 4.3. To further enhance consistency, we apply per-step noise and prediction consistency guidance [89], both spatially modulated by the reliability weight r . First, at each denoising step $t \in \{0, \dots, T-1\}$, we compute a reliability-weighted reference noise from the previous frame:

$$\eta_{\text{ref}}^{i-1} = D \odot \mathcal{W}(\eta_t^{i-1}, \mathbf{F}_{i-1 \rightarrow i}) + (1 - D) \odot \eta_t^{i-1}, \quad (4.6)$$

and replace the current stochastic noise as:

$$\tilde{\eta}_t^i = (1 - M_{\text{occ}}) \odot \eta_{\text{ref}}^{i-1} + M_{\text{occ}} \odot \eta_{\text{rand}}, \quad (4.7)$$

where non-occluded regions adopt the reference noise weighted by event density, and occluded regions are filled with fresh random noise η_{rand} . Second, at each denoising step $t \in \{T/2, \dots, T-1\}$, we impose gradient-based consistency on the intermediate clean latent $\hat{x}_0^i$ estimated by diffusion UNet toward the fully denoised latent of the previous frame z_0^{i-1} by minimizing:

$$\mathcal{L}_{\text{PCG}} = \|(1 - M_{\text{occ}}) \odot (\hat{x}_0^i - \hat{z}_{\text{ref}}^{i-1})\|_1, \quad (4.8)$$

Table 4.1: Quantitative evaluation on REVD and EVRB datasets. The best and second-best results are highlighted in **bold** and underlined, respectively. Our method significantly improves the temporal consistency (tOF, tLP, WE) of the pre-trained IDM (DeblurDiff) while preserving its high spatial restoration quality (PSNR, SSIM).

Method	REVD Dataset					EVRB Dataset				
	PSNR↑	SSIM↑	tOF↓	tLP↓	WE↓	PSNR↑	SSIM↑	tOF↓	tLP↓	WE↓
MAENet	25.41	0.813	1.043	0.0233	0.0459	25.46	0.791	1.657	0.0249	<u>0.0413</u>
DeblurDiff	<u>26.53</u>	<u>0.822</u>	1.213	0.0273	0.0531	27.34	0.814	1.702	0.0293	0.0437
DeblurDiff + Ours	26.58	0.825	<u>1.122</u>	<u>0.0235</u>	<u>0.0476</u>	<u>27.28</u>	<u>0.812</u>	<u>1.689</u>	<u>0.0264</u>	0.0412

where $\hat{\mathbf{z}}_{\text{ref}}^{i-1} = D \odot \mathcal{W}(\mathbf{z}_0^{i-1}, \mathbf{F}_{i-1 \rightarrow i}) + (1 - D) \odot \mathbf{z}_0^{i-1}$. Next, we update via gradient descent:

$$\tilde{\mathbf{x}}_0^i \leftarrow \hat{\mathbf{x}}_0^i - \lambda \cdot \frac{1}{\mu(t)} \nabla \mathcal{L}_{\text{PCG}}, \quad (4.9)$$

where $\mu(t)$ is a step-dependent normalization factor [92]. This guidance is applied only for $t \geq T/2$, as early-step estimates are unreliably noisy [89].

4.4 Experiments

4.4.1 Experimental Setups

Baseline. To demonstrate the effectiveness of our zero-shot plug-and-play framework, we adopt the state-of-the-art pre-trained Image Diffusion Model (IDM), DeblurDiff [84], as our primary baseline. Consistent with our zero-shot setting, all weights of the pre-trained IDM remain strictly frozen during the inference. We also include SOTA regression-based method MAENet [5] for comparison.

Datasets. We evaluate on two real-world event-based video deblurring benchmarks: EVRB [79] and REVD [80]. These datasets provide real-world motion blur paired with high-temporal-resolution event streams, which aligns with our motivation to validate event-guided motion reliability under dynamic scenes.

Evaluation Metrics. We measure spatial restoration quality using PSNR and SSIM [49]. To assess our primary objective—temporal consistency—we employ tOF, tLP [51], and Warping Error (WE) [89].

4.4.2 Experimental Results

Quantitative Results. Table 4.1 shows that our method effectively balances spatial quality and temporal consistency. Compared to the zero-shot IDM baseline (DeblurDiff), our approach reduces temporal artifacts, leading to lower tOF, tLP, and WE scores across both the REVD and EVRB datasets. Concurrently, it maintains the spatial fidelity (PSNR and SSIM) of the pre-trained diffusion model. As a result, our framework achieves temporal stability comparable to the regression-based MAENet while preserving the visual quality provided by generative priors.

Qualitative Results. Visual comparisons on consecutive frames (See Fig. 4.4, 4.5) demonstrate the effectiveness of our approach in reducing temporal artifacts. As shown in the visual examples, the regression-based MAENet yields temporally stable outputs but tends to produce overly smooth results, losing fine structural details. The zero-shot IDM baseline, DeblurDiff, generates sharp and realistic textures. However, it exhibits noticeable temporal inconsistency; the generated background

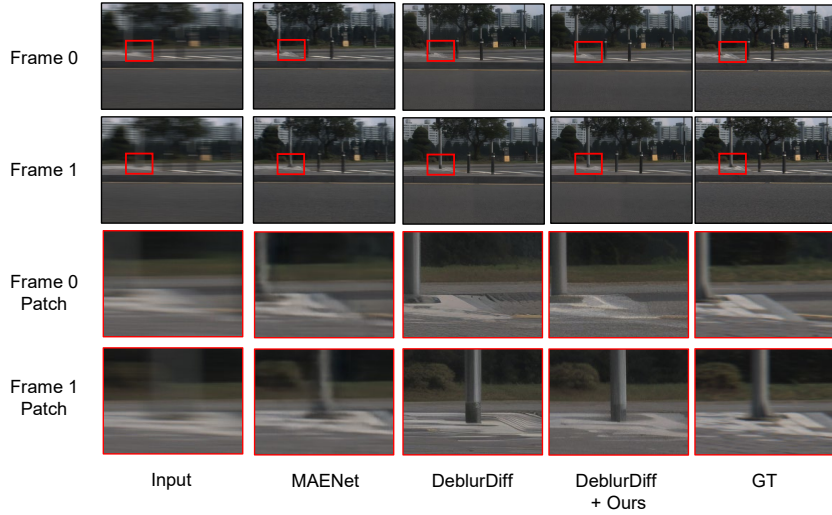


Figure 4.4: Qualitative comparison across consecutive frames. MAENet produces overly smoothed results, and DeblurDiff suffers from severe temporal inconsistency. In contrast, our method generates both temporally consistent while preserving its deblurring quality.



Figure 4.5: Additional qualitative comparison.

patterns—such as the tactile paving on the ground and the texture of the rocks—shift or change between adjacent frames. By incorporating our spatially adaptive guidance, our method preserves the high spatial quality of the generative prior while ensuring consistent structural generation across frames, effectively mitigating flickering.

Ablation Studies. Table 4.2 investigates the individual contributions of our proposed spatial guidance components: Cross-frame Attention (CA), Noise Consistency Guidance (NCG), and Prediction Consistency Guidance (PCG). Starting from the zero-shot IDM baseline, applying any single component independently reduces the Warping Error (WE), indicating improved temporal consistency. CA provides the most notable individual improvement by effectively propagating structural context. Furthermore, integrating all three mechanisms yields the lowest WE score, demonstrating that modulating these constraints at complementary levels of the diffusion process minimizes temporal artifacts more effectively than using any single component.

Table 4.2: Ablation study on the effectiveness of different guidance components evaluated on the REVD dataset. CA: Cross-frame Attention, NCG: Noise Consistency Guidance, PCG: Prediction Consistency Guidance.

CA	NCG	PCG	WE↓
			0.0531
✓			0.0488
	✓		0.0523
		✓	0.0517
✓	✓	✓	0.0476

4.5 Conclusion

In this study, we presented a plug-and-play, zero-shot video deblurring framework that enhances the temporal consistency of pre-trained image diffusion models using event cameras. Recognizing the unreliability of explicit motion estimation under severe blur, we utilized event density as a spatial reliability score. This allowed us to adaptively modulate cross-frame attention, noise consistency, and prediction consistency based on local motion confidence. Experimental results indicate that our spatially adaptive guidance effectively balances temporal stability and spatial restoration quality without requiring any fine-tuning. We hope our approach offers a practical and efficient step toward reliable, training-free video restoration in dynamic real-world scenarios.

Chapter 5. Conclusion

This dissertation has explored how to achieve robust image and video restoration by carefully managing the inherent physical limitations of event cameras. Rather than treating the event modality as a flawless guide, we proposed practical frameworks for real-world frame restoration. Specifically, we demonstrated how events can serve as reliable guidance when their temporal ambiguity and spatial sparsity are properly controlled. The primary contributions are summarized as follows:

- **Robust Restoration under Blind and Dynamic Exposures (Chapter 2):** We introduced a joint video frame interpolation and motion deblurring framework designed to handle blind exposure scenarios. By modeling temporal constraints through an adaptive feature blending strategy, the proposed method reconstructs sharp target frames even when the target timestamp deviates significantly from the exposure period, demonstrating improved generalization over prior methods.
- **Overcoming Domain Shifts via Motion-Content Decoupling (Chapter 3):** We tackled the domain shift problem in motion deblurring by presenting an event-guided blur synthesis pipeline. By utilizing the spatiotemporal precision of events to decouple spatial content from motion, our approach addresses the limitations of existing self-reblurring methods that rely solely on intrinsic motion. By transferring both intrinsic and extrinsic motions, we synthesized diverse and realistic training pairs, enabling the effective calibration of pre-trained models to target domains.
- **Zero-shot Temporal Consistency in Generative Models (Chapter 4):** We proposed a training-free framework to mitigate the temporal inconsistency of pre-trained image diffusion models when applied to video processing. By employing event density as a spatial reliability metric, our method adaptively modulates the generative prior. This approach balances high-quality spatial restoration with temporal stability, effectively enhancing temporal consistency without the need for computationally expensive video diffusion models.

While this dissertation establishes a robust foundation for real-world low-level image and video restoration, the unique spatiotemporal properties of event cameras present immense opportunities for broader high-level vision and multi-modal AI. One promising direction for future research is the integration of the event modality into large-scale Vision-Language Models (VLMs) and foundation models. Since current VLMs predominantly rely on discrete RGB frames, they often struggle with fine-grained temporal reasoning in highly dynamic or extreme lighting scenarios. Introducing events as a complementary modality could provide these models with continuous, high-dynamic-range motion cues, significantly enhancing complex video understanding and dynamic scene reasoning.

Furthermore, fully exploiting the sparse, asynchronous nature of events can drive the development of highly efficient, event-aware perception systems for resource-constrained edge devices and embodied robotics. Ultimately, expanding event-guided frameworks from low-level restoration to high-level multi-modal reasoning will pave the way for next-generation AI systems capable of robustly perceiving and interacting with the unpredictable real world.

Bibliography

- [1] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong, “Event-based blurry frame interpolation under blind exposure”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [2] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring”, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [3] Po Sheng Liu, Fu Jen Tsai, Yan Tsung Peng, Chung Chi Tsai, Chia Wen Lin, and Yen Yu Lin, “Meta transferring for deblurring”, in *British Machine Vision Conference Proceedings (BMVC)*, 2022.
- [4] Jin-Ting He, Fu-Jen Tsai, Jia-Hao Wu, Yan-Tsung Peng, Chung-Chi Tsai, Chia-Wen Lin, and Yen-Yu Lin, “Domain-adaptive video deblurring via test-time blurring”, in *European Conference on Computer Vision (ECCV)*, 2024.
- [5] Zhijing Sun, Xueyang Fu, Longzhuo Huang, Aiping Liu, and Zheng-Jun Zha, “Motion aware event representation-driven image deblurring”, in *European Conference on Computer Vision (ECCV)*. Springer, 2024.
- [6] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel, “Benchmarking robustness in object detection: Autonomous driving when winter is coming”, *arXiv preprint arXiv:1907.07484*, 2019.
- [7] Jiong Dong, Kaoru Ota, and Mianxiong Dong, “Video frame interpolation: A comprehensive survey”, *ACM Transactions on Multimedia Computing, Communications and Applications (TOMM)*, 2023.
- [8] Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li, “Deep image deblurring: A survey”, *International Journal of Computer Vision (IJCV)*, 2022.
- [9] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al., “Event-based vision: A survey”, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.
- [10] Bharatesh Chakravarthi, Aayush Atul Verma, Kostas Daniilidis, Cornelia Fermüller, and Yezhou Yang, “Recent event camera innovations: A survey”, in *European conference on computer vision (ECCV)*, 2024.
- [11] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza, “Time lens: Event-based video frame interpolation”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [12] Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, and Jianxing Liao, “Timereplayer: Unlocking the potential of event cameras

- for video interpolation”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [13] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza, “Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
 - [14] 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 *European Conference on Computer Vision (ECCV)*, 2020.
 - [15] Xiang Zhang and Lei Yu, “Unifying motion deblurring and frame interpolation with events”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
 - [16] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhong Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool, “Event-based frame interpolation with ad-hoc deblurring”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
 - [17] Blain Brown, *Cinematography: theory and practice: image making for cinematographers and directors*, Taylor & Francis, 2016.
 - [18] Taewoo Kim, Jeongmin Lee, Lin Wang, and Kuk-Jin Yoon, “Event-guided deblurring of unknown exposure time videos”, in *European Conference on Computer Vision (ECCV)*, 2022.
 - [19] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou, “Real-time intermediate flow estimation for video frame interpolation”, in *European Conference on Computer Vision (ECCV)*, 2022.
 - [20] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang, “Ifnet: Intermediate feature refine network for efficient frame interpolation”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
 - [21] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai, “Bringing a blurry frame alive at high frame-rate with an event camera”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
 - [22] Dawit Mureja Argaw and In So Kweon, “Long-term video frame interpolation via feature propagation”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
 - [23] Simon Niklaus and Feng Liu, “Softmax splatting for video frame interpolation”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
 - [24] Ping Hu, Simon Niklaus, Stan Sclaroff, and Kate Saenko, “Many-to-many splatting for efficient video frame interpolation”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
 - [25] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless, “Film: Frame interpolation for large motion”, in *European Conference on Computer Vision (ECCV)*, 2022.

- [26] Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee, “Adacof: Adaptive collaboration of flows for video frame interpolation”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [27] Simon Niklaus, Long Mai, and Feng Liu, “Video frame interpolation via adaptive convolution”, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [28] Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia, “Video frame interpolation with transformer”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [29] Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, and Ming-Hsuan Yang, “Video frame interpolation transformer”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [30] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee, “Channel attention is all you need for video frame interpolation”, in *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.
- [31] Zhihao Shi, Xiaohong Liu, Kangdi Shi, Linhui Dai, and Jun Chen, “Video frame interpolation via generalized deformable convolution”, *IEEE Transactions on Multimedia (TMM)*, vol. 24, pp. 426–439, 2021.
- [32] Maitreya Suin and AN Rajagopalan, “Gated spatio-temporal attention-guided video deblurring”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [33] Bo Ji and Angela Yao, “Multi-scale memory-based video deblurring”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [34] Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu, “Learning event-based motion deblurring”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [35] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool, “Event-based fusion for motion deblurring with cross-modal attention”, in *European Conference on Computer Vision (ECCV)*, 2022.
- [36] Kang Chen and Lei Yu, “Motion deblur by learning residual from events”, *IEEE Transactions on Multimedia (TMM)*, 2024.
- [37] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun, “Simple baselines for image restoration”, in *European conference on computer vision (ECCV)*, 2022, pp. 17–33.
- [38] Meiguang Jin, Zhe Hu, and Paolo Favaro, “Learning to extract flawless slow motion from blurry videos”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [39] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao, “Blurry video frame interpolation”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [40] Youjian Zhang, Chaoyue Wang, and Dacheng Tao, “Video frame interpolation without temporal priors”, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

- [41] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation”, in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.
- [43] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza, “Event-based vision meets deep learning on steering prediction for self-driving cars”, in *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 5419–5427.
- [44] Zhijian Wu, Jun Li, Chang Xu, Dingjiang Huang, and Steven CH Hoi, “Run: Rethinking the unet architecture for efficient image restoration”, *IEEE Transactions on Multimedia (TMM)*, 2024.
- [45] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution”, in *European Conference on Computer Vision (ECCV)*, 2016.
- [46] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza, “Esim: an open event camera simulator”, in *Conference on Robot Learning (CoRL)*, 2018.
- [47] Ilya Loshchilov and Frank Hutter, “Decoupled weight decay regularization”, *arXiv preprint arXiv:1711.05101*, 2017.
- [48] Ilya Loshchilov and Frank Hutter, “Sgdr: Stochastic gradient descent with warm restarts”, *arXiv preprint arXiv:1608.03983*, 2016.
- [49] 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 (TIP)*, 2004.
- [50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, “The unreasonable effectiveness of deep features as a perceptual metric”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [51] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey, “Learning temporal coherence via self-supervision for gan-based video generation”, *ACM Transactions on Graphics (TOG)*, 2020.
- [52] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, “Making a “completely blind” image quality analyzer”, *IEEE Signal processing letters (SPL)*, 2012.
- [53] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization”, in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.
- [54] Yunfan Lu, Zipeng Wang, Minjie Liu, Hongjian Wang, and Lin Wang, “Learning spatial-temporal implicit neural representations for event-guided video super-resolution”, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [55] Xiang Zhang, Lei Yu, Wen Yang, Jianzhuang Liu, and Gui-Song Xia, “Generalizing event-based motion deblurring in real-world scenarios”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2023, pp. 10734–10744.

- [56] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko, “Rethinking coarse-to-fine approach in single image deblurring”, in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4641–4650.
- [57] 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)*, 2023.
- [58] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck, “A 128×128 120 db $15\mu s$ latency asynchronous temporal contrast vision sensor”, *IEEE journal of solid-state circuits*, 2008.
- [59] Wen Yang, Jinjian Wu, Leida Li, Weisheng Dong, and Guangming Shi, “Event-based motion deblurring with modality-aware decomposition and recomposition”, in *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 2023.
- [60] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho, “Real-world blur dataset for learning and benchmarking deblurring algorithms”, in *European conference on computer vision (ECCV)*, 2020, pp. 184–201.
- [61] Zhixiang Chi, Yang Wang, Yuanhao Yu, and Jin Tang, “Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning”, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021.
- [62] Seungjun Nah, Sanghyun Son, Jaerin Lee, and Kyoung Mu Lee, “Clean images are hard to reblur: Exploiting the ill-posed inverse task for dynamic scene deblurring”, in *International Conference on Learning Representations (ICLR)*, 2022.
- [63] Jia-Hao Wu, Fu-Jen Tsai, Yan-Tsung Peng, Chung-Chi Tsai, Chia-Wen Lin, and Yen-Yu Lin, “Id-blau: Image deblurring by implicit diffusion-based reblurring augmentation”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [64] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan, “Spatially-attentive patch-hierarchical network for adaptive motion deblurring”, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020.
- [65] Fu-Jen Tsai, Yan-Tsung Peng, Chung-Chi Tsai, Yen-Yu Lin, and Chia-Wen Lin, “Banet: A blur-aware attention network for dynamic scene deblurring”, *IEEE Transactions on Image Processing (TIP)*, 2022.
- [66] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin, “Stripformer: Strip transformer for fast image deblurring”, in *European conference on computer vision (ECCV)*, 2022.
- [67] Xin Gao, Tianheng Qiu, Xinyu Zhang, Hanlin Bai, Kang Liu, Xuan Huang, Hu Wei, Guoying Zhang, and Huaping Liu, “Efficient multi-scale network with learnable discrete wavelet transform for blind motion deblurring”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [68] 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 (NeurIPS)*, 2023.

- [69] Yanan Zhu, Jiaqiu Ai, Le Wu, Dan Guo, Wei Jia, and Richang Hong, “An active multi-target domain adaptation strategy: Progressive class prototype rectification”, *IEEE Transactions on Multimedia (TMM)*, 2024.
- [70] Yuwu Lu, Haoyu Huang, Biqing Zeng, Zhihui Lai, and Xuelong Li, “Multi-source and multi-target domain adaptation based on dynamic generator with attention”, *IEEE Transactions on Multimedia (TMM)*, 2024.
- [71] Wuyang Li, Xinyu Liu, and Yixuan Yuan, “Sigma: Semantic-complete graph matching for domain adaptive object detection”, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022.
- [72] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda, “Cross-domain adaptive teacher for object detection”, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022.
- [73] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool, “Mic: Masked image consistency for context-enhanced domain adaptation”, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023.
- [74] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T Tan, “Cat: Exploiting inter-class dynamics for domain adaptive object detection”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [75] Youngho Kim, Hoonhee Cho, and Kuk-Jin Yoon, “From sharp to blur: Unsupervised domain adaptation for 2d human pose estimation under extreme motion blur using event cameras”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.
- [76] Ye-Chan Kim, SeungJu Cha, Si-Woo Kim, Taewhan Kim, and Dong-Jin Kim, “Sida: Synthetic image driven zero-shot domain adaptation”, in *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025.
- [77] Jiankun Zhu, Sicheng Zhao, Lulu Tian, Jing Jiang, Xi Chen, and Hongxun Yao, “Emotion in a bottle: Information bottleneck guided disentanglement for emotion domain adaptation”, in *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, 2025.
- [78] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza, “E-raft: Dense optical flow from event cameras”, in *2021 International Conference on 3D Vision (3DV)*, 2021.
- [79] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon, “Cmta: Cross-modal temporal alignment for event-guided video deblurring”, in *European Conference on Computer Vision (ECCV)*, 2024.
- [80] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon, “Frequency-aware event-based video deblurring for real-world motion blur”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [81] Jaesung Rim, Geonung Kim, Jungeon Kim, Junyong Lee, Seungyong Lee, and Sunghyun Cho, “Realistic blur synthesis for learning image deblurring”, in *European Conference on Computer Vision (ECCV)*, 2022.

- [82] Diederik Kinga, Jimmy Ba Adam, et al., “A method for stochastic optimization”, in *International conference on learning representations (ICLR)*, 2015.
- [83] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang, “Quadratic video interpolation”, *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [84] Lingshun Kong, Jiawei Zhang, Dongqing Zou, Fu Lee Wang, Jimmy S Ren, Xiaohe Wu, Jiangxin Dong, and Jinshan Pan, “Deblurdiff: Real-world image deblurring with generative diffusion models”, *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.
- [85] Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye, “Parallel diffusion models of operator and image for blind inverse problems”, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023.
- [86] Shitong Shao, Lichen Bai, Pengfei Wan, James Kwok, and Zeke Xie, “Efficient video diffusion models: Advancements and challenges”, *arXiv preprint arXiv:2604.15911*, 2026.
- [87] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi, “Text2video-zero: Text-to-image diffusion models are zero-shot video generators”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [88] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou, “Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation”, in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2023.
- [89] Cong Cao, Huanjing Yue, Xin Liu, and Jingyu Yang, “Zero-shot video restoration and enhancement using pre-trained image diffusion model”, in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.
- [90] Jonathan Ho, Ajay Jain, and Pieter Abbeel, “Denoising diffusion probabilistic models”, *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [91] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song, “Denoising diffusion restoration models”, *Advances in neural information processing systems (NeurIPS)*, 2022.
- [92] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye, “Diffusion posterior sampling for general noisy inverse problems”, *arXiv preprint arXiv:2209.14687*, 2022.
- [93] Xiaoyang Liu, Zhengyan Zhou, Zihang Xu, Jiezhong Cao, Zheng Chen, and Yulun Zhang, “Fidediff: Efficient diffusion model for high-fidelity image motion deblurring”, *arXiv preprint arXiv:2510.01641*, 2025.
- [94] Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Chen Change Loy, and Lu Jiang, “Seedvr: Seeding infinity in diffusion transformer towards generic video restoration”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [95] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy, “Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution”, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2024.

- [96] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang, “Learning blind video temporal consistency”, in *Proceedings of the European conference on computer vision (ECCV)*, 2018.
- [97] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or, “Prompt-to-prompt image editing with cross attention control”, *arXiv preprint arXiv:2208.01626*, 2022.