EXCHANGE, NOT GATING: EVENT-GUIDED DEBLURRING AND DERAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Event cameras record brightness changes asynchronously at high temporal resolution, making them a natural guide for RGB restoration tasks such as motion deblurring and image deraining. Existing methods, however, still couple the two modalities through three heuristic choices: they process event voxels with blocks designed for images, control event injection with hand-designed activity masks and unconstrained gates, and decode fused features without directional guidance. We replace these choices with three complementary mechanisms. VIBE provides event-native processing by separating polarities and conditioning the event stream on the mass, signed mass, temporal centroid and temporal spread of each pixel's event histogram, thereby removing the baseline's hand-designed blur-and-mask pipeline. COEX replaces independent gates with a per-channel 2×2 mixer projected toward the doubly stochastic manifold, encouraging convex cross-modal mixing while approximately preserving the sum of the two inputs. COMPASS predicts a per-pixel distribution over five fixed shifts from event features and their moment condition, realigning fused features without regressing offsets. The resulting single-frame model achieves the best event-guided results on GoPro, HighREV and NTURain: on GoPro, it reaches 37.72 dB compared with 37.10 dB from the official TRM-UNet weights, while a larger configuration reaches 37.97 dB.

1 INTRODUCTION

Motion deblurring and image deraining arise from the same limitation of a finite exposure: transient changes are integrated into a frame, obscuring their trajectories. Event cameras retain part of this lost temporal information by responding asynchronously to brightness changes at microsecond resolution, making them a natural guide for both restoration tasks (Yang et al., 2023; 2025; Sun et al., 2025a; Zhang et al., 2025; Sun et al., 2023b). Frames and events are therefore complementary—the former are dense but temporally collapsed, whereas the latter are sparse but temporally precise—which makes effective cross-modal fusion the central design challenge.

Current event-guided networks largely address this challenge through three heuristic design choices.

(i) Event voxels are processed like images. Once constructed, a voxel grid is typically passed through the same convolutional or state-space blocks used for RGB features, with a hand-designed mask providing the main event-specific operation. This treatment overlooks the structure of the voxel itself: its sign encodes the direction of brightness change, while its distribution across temporal bins describes when that change occurs. The common mask pipeline further compresses the voxel into a blurred activity map before learning begins, making it invariant to both polarity allocation and temporal-bin order. This information loss is substantial: on the GoPro (Nah et al., 2017), HighREV (Sun et al., 2023a) and NTURain (Chen et al., 2018) validation sets, 16.0%, 18.2% and 41.4% of pixels contain both polarities, yet their net signed evidence is only 27%, 40% and 27% of the total activity read by the mask.

(ii) Fusion is a gate. The dominant design first admits event evidence through an activity mask and then reweights the two streams with gates (Yang et al., 2023; 2025). At the sites considered in this work, the two gains are predicted independently and need not sum to one; fusion can therefore rescale either modality instead of explicitly redistributing information between them. Sum-to-one mixing does appear in prior work, but only at isolated sites and with weights predicted from globally pooled features. No single constraint governs the interaction at every fusion site.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(iii) The readout is direction-agnostic. Restoration decoders repeatedly fuse, upsample and refine features, but conventional readouts do not explicitly model the direction along which motion or rain has displaced local structure. Deformable convolution addresses this issue by regressing continuous offsets, at the cost of an additional prediction branch and potentially unstable offset estimation.

We replace these three heuristics with targeted mechanisms while keeping the backbone units, training recipe and data pipeline fixed, enabling controlled attribution of the resulting gains.

Event-specific blocks (VIBE). VIBE treats each voxel pixel as a signed temporal histogram rather than an image channel. It separates positive and negative events and conditions both paths on four statistics: total mass, signed mass, temporal centroid and temporal spread. A learned affine maps these statistics to per-channel scales and shifts. Unlike the baseline mask, which retains only blurred activity, VIBE preserves polarity and temporal information without hand-designed masking.

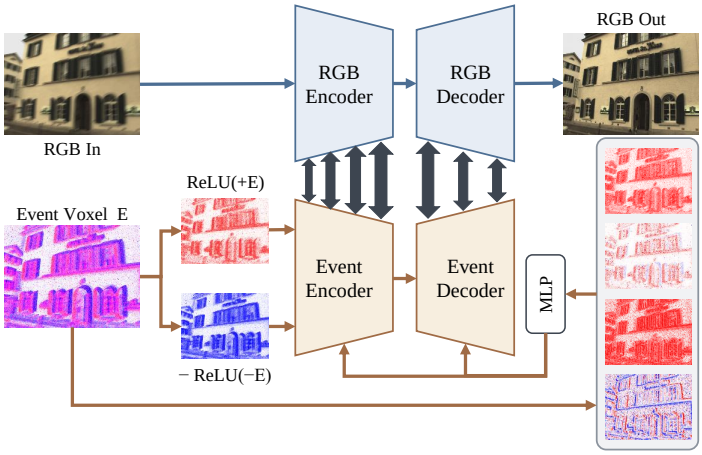

Figure 1: **The backbone and the sites we change.** Polarity is split explicitly and the streams are exchanged at every fusion site.

Conserved exchange (COEX). COEX formulates every fusion site as a per-channel exchange between the two streams. A lightweight predictor produces a 2×2 mixing matrix from a joint cross-modal descriptor, and iterative normalisation projects this matrix toward the doubly stochastic manifold. The resulting coefficients approximately form convex combinations and preserve the sum of the two inputs within the mixer, discouraging coefficient-sum amplification. The projection introduces no learned parameters beyond the matrix predictor, and its coefficients remain interpretable as retained and exchanged fractions.

Direction-guided readout (COMPASS). COMPASS aggregates five fixed views of the fused feature at each decoder scale: the centre and its four axial neighbours. A shared pointwise MLP maps the event feature $\mathbf{Y}$ and condition $\mathbf{C}$ to a per-pixel distribution over these views. Because the sampling locations remain fixed, COMPASS selects a local direction without regressing continuous offsets.

Together, the three mechanisms address event encoding, cross-modal interaction and decoder readout (Fig. 1). The complete model reaches 37.72 dB on GoPro, compared with 37.10 dB from the official TRM-UNet weights, while its larger configuration reaches 37.97 dB. It also achieves 37.03 dB on HighREV and 46.38 dB on NTURain; we report these results together with MACs and parameter counts.

2 Related Work

Event-guided restoration. Event-guided deblurring is dominated by dual-stream designs in which an event branch feeds a frame restoration network: cascaded variants re-inject event features at successive stages (Sun et al., 2022; Guo et al., 2026), while persistent variants keep two streams and combine them with cross-modal attention or gates (Sun et al., 2024; Yang et al., 2023; 2024; Xu et al., 2025; Yang et al., 2025; Xiao et al., 2026; Sun et al., 2025a; Fan et al., 2026). Event-guided deraining is predominantly temporal: recurrent or multi-frame networks consume a window of frames together with event voxels (Wang et al., 2023; Sun et al., 2023b; Zhang et al., 2025), and recent designs add spiking dynamics or semi-supervised training (Sun et al., 2025b; Wu et al., 2024; Sun et al., 2026; Jiang et al., 2026). What these works share is a fusion rule that is a *decision*: a mask or a gate asks how much event evidence to admit, per location or per channel. A decision carries no constraint. At the sites we replace, the two gains are predicted independently and may amplify a modality; and the mask that decides *where* events enter is hand-designed from the voxel

and uses the same fixed functional form for every sample. In AHDINet (Yang et al., 2025) it is $M = \mathrm{clip}\big(G_\sigma * (G_\sigma * \sum_t |\mathbf{E}_t|), 0, 1\big)$, a function of the activity alone, hence exactly invariant to the polarity split and to any permutation of the temporal bins: on the GoPro (Nah et al., 2017), HighREV (Sun et al., 2023a) and NTURain (Chen et al., 2018) validation sets the polarity contrast $|P - N|/(P + N)$ at pixels carrying both polarities averages only 0.27, 0.40 and 0.27, while the clamp saturates 27%, 49% and 65% of pixels at 1. Conserved, sum-to-one mixing also appears in prior work: the fusion at two of AHDINet's four levels yields per-channel weights $(w, 1 - w)$, but from pooled features alone. Our exchange (COEX) replaces the decision by a projection: the mixing matrix is projected toward the doubly stochastic manifold at all seven fusion sites, its descriptor also observes the inter-modal disagreement $|\mathbf{X} - \mathbf{Y}|$, and the resulting constraint discourages coefficient-sum amplification inside the mixer.

Event representations. Because an event camera reports signed brightness changes, a stream must first become a tensor; standard choices are voxel grids, time surfaces and event frames (Sun et al., 2024; Yang et al., 2023), and deraining pipelines usually follow the released ten-channel rainy voxels (Zhang et al., 2025). The representation carries modality-specific structure: polarity is a sign, and the distribution of a voxel over its bins is a temporal statistic. In the networks above, however, the resulting tensor is processed by the same convolution and state-space blocks as the frame, and the hand-designed mask is the only event-specific step in the whole pipeline. Our blocks (VIBE) keep polarity as two paths and condition the event features on the mass, the signed mass, the temporal centroid and the temporal spread of that histogram, so the hand-designed mask is removed rather than re-tuned.

Alignment and readout. Sub-pixel structure along motion does not survive decoder upsampling, and the standard remedies are to regress sampling offsets, as in deformable convolution (Dai et al., 2017), or to maintain explicit temporal state, as in video derainers (Wu et al., 2024; Sun et al., 2025b; 2026). Offset regression is flexible but needs an extra prediction branch and is known to be unstable when the offset field is large; temporal state is effective but presumes a sequence at inference. Our readout (COMPASS) takes a middle route: a pointwise MLP of the event feature and its moment condition predicts a per-pixel distribution over five fixed shifts, choosing a direction without regressing a position and without carrying state across frames.

Efficient restoration backbones. Linear-cost sequence models have become competitive backbones for restoration (Guo et al., 2025; Lin et al., 2025b; Li et al., 2025), and convolution-based designs such as ConvIR (Cui et al., 2024) remain strong at low cost. We adopt such units unchanged: the contribution here is not a new backbone, but the way event information is encoded, exchanged and read out within it. Our controlled comparisons therefore share backbone widths and training settings while varying the proposed mechanisms.

3 METHOD

Our starting point is a standard dual-stream architecture comprising frame and event encoders, a shared state-space unit at each stage, and a decoder that progressively combines the two streams (Fan et al., 2026; Sun et al., 2025a). We retain its backbone units, stage widths and training recipe, and modify only event encoding, cross-modal fusion and decoder readout. As illustrated in Fig. 1, VIBE introduces polarity-aware event processing at every level, COEX exchanges information between the two streams at all fusion sites, and COMPASS performs direction-guided aggregation at the three decoder scales. The three modules follow a common design principle: event measurements should determine *what* temporal evidence is available, cross-modal interaction should determine *how* that evidence is shared, and the decoder should determine *where* it should be read. This decomposition assigns a distinct role to each mechanism and avoids asking a single attention block to simultaneously encode events, estimate their reliability and align the reconstruction. It also yields direct controls for the ablation study: event statistics can be reduced to activity, the exchange projection can be removed, and spatial readout weights can be made global without changing the surrounding backbone. The mechanisms remain connected without collapsing these roles: VIBE supplies event features and moment conditions, COEX exchanges channels without spatial resampling, and COMPASS selects local views of the fused feature. COEX therefore adapts channel contributions from cross-modal agreement, whereas COMPASS varies direction from pixel to pixel using only five fixed decoder taps.

3.1 EVENT-NATIVE BLOCKS

The baseline prepares its event stream with a fixed mask pipeline: $M = \mathrm{clip}\big(G_\sigma * (G_\sigma * \sum_t |\mathbf{E}_t|), 0, 1\big)$. Because this mask depends only on activity $\sum_t |\mathbf{E}_t|$, it is invariant to both the allocation of activity across polarities and any permutation of the temporal bins. It therefore cannot convey polarity balance or event timing to the branch. VIBE instead represents each voxel pixel as a signed temporal histogram, using the sign to retain the direction of brightness change and the bin distribution to describe its timing.

The voxel $\mathbf{E}$ is split into $\mathrm{ReLU}(+\mathbf{E})$ and $\mathrm{ReLU}(-\mathbf{E})$. With bin times t placed uniformly in $[-1, 1]$, each pixel is summarised by

$$\sum_t |\mathbf{E}_t|, \quad \sum_t \mathbf{E}_t, \quad \bar{t} = \frac{\sum_t |\mathbf{E}_t|\, t}{\sum_t |\mathbf{E}_t| + \epsilon}, \quad \frac{\sum_t |\mathbf{E}_t|(t - \bar{t})^2}{\sum_t |\mathbf{E}_t| + \epsilon}, \tag{1}$$

where $\epsilon = 10^{-6}$, matching the implementation. In that order, these are the total mass, the signed mass, the temporal centroid, and the temporal spread; the latter two are zero at pixels with no events. These statistics are complementary: total mass measures event support, signed mass restores polarity imbalance, the centroid locates activity within the exposure, and the spread distinguishes compact from temporally distributed responses. Computing them per pixel preserves spatial variation that global pooling would discard. These four maps form the condition $\mathbf{C}$. A learned affine maps $\mathbf{C}$ to per-channel scales and shifts for the two polarity paths, which are then processed by the unchanged backbone units. Thus, VIBE replaces the hand-designed mask rather than re-tuning it, allowing the event branch to learn from measured polarity and timing statistics. We apply VIBE at all seven stages—four encoder stages and three decoder stages—and report its additional computational cost in Sec. 4.

3.2 CONSERVED EXCHANGE

At each fusion site, the network must determine the contribution of frame and event features. The baseline combines a spatial activity mask with channel gates, whose independently predicted gains need not sum to one. For example, AHDINet scales the event branch by $(1 + M)w$ while leaving the frame branch unchanged, and its decoder fusion applies two independent sigmoids. Although AHDINet uses weights $(w, 1 - w)$ at two fusion levels, those weights are predicted from pooled features and do not impose a common constraint across all sites. COEX instead treats fusion as an exchange governed by an approximate sum-preserving constraint.

For each channel, we predict a 2×2 matrix $\mathbf{M}$ from a globally pooled descriptor containing $\mathbf{X}$, $\mathbf{Y}$ and their disagreement $|\mathbf{X} - \mathbf{Y}|$. Two 1×1 convolutions map this descriptor to the matrix coefficients, and the mixed features are

$$\tilde{\mathbf{X}}_i = \mathbf{M}_{00}\, \mathbf{X}_i + \mathbf{M}_{01}\, \mathbf{Y}_i, \qquad \tilde{\mathbf{Y}}_i = \mathbf{M}_{10}\, \mathbf{X}_i + \mathbf{M}_{11}\, \mathbf{Y}_i, \tag{2}$$

We exponentiate the predicted logits and apply alternating row and column normalisation for $T = 5$ rounds, projecting $\mathbf{M}$ toward the doubly stochastic manifold. Because the projection is finite, its row and column sums are close to, rather than exactly, one. Each output is therefore approximately a convex combination of the two modalities, with $\tilde{\mathbf{X}}_i + \tilde{\mathbf{Y}}_i \approx \mathbf{X}_i + \mathbf{Y}_i$ within the mixing operator. The entries remain interpretable as mixing coefficients: $\mathbf{M}_{00}$ is the frame fraction retained by the frame output, $\mathbf{M}_{01}$ is the event fraction transferred to it, and the second row has the symmetric interpretation. The projection requires neither additional learned parameters nor an auxiliary loss. Including both modalities and their absolute disagreement serves different purposes: the pooled frame and event descriptors summarise the content available in each stream, while $|\mathbf{X} - \mathbf{Y}|$ exposes channels for which the two streams disagree. The predictor can therefore adapt the exchange to cross-modal consistency rather than estimating two unrelated gains. Because the matrix is predicted per channel, reliable event channels can contribute strongly without forcing the same exchange ratio on appearance channels that are already well explained by the frame.

A doubly stochastic 2×2 matrix with a zero off-diagonal entry is necessarily the identity. Thus, under this constraint, conservation and reciprocal exchange are coupled: a one-way exchange is a different operator rather than a weaker form of COEX. We use the same mixer at all seven fusion sites—four in the encoder and three in the decoder—while adapting its conditioning to each stage.

3.3 DIRECTION-GUIDED READOUT

The decoder contains three fusion sites separated by upsampling stages. Conventional upsampling does not distinguish which neighbouring feature is most relevant along the local motion or rain direction. Deformable convolution addresses this limitation by regressing sampling offsets, but requires an additional prediction branch and can become unstable for large offsets; temporal models instead require a frame sequence at inference. COMPASS keeps the sampling positions fixed and learns only their aggregation weights.

At each of the three sites, let $\mathbf{F}$ be the fused feature, $\mathbf{Y}$ the event feature, and $\mathbf{C}$ the condition in equation 1 at that resolution. The weights are a distribution over the centre and the four axis neighbours, predicted from the event side,

$$\mathbf{w} = \mathrm{softmax}\big(\mathrm{MLP}([\mathbf{Y}, \mathbf{C}])\big) \in \mathbb{R}^{H\times W\times 5}, \qquad \mathbf{F}'(p) = \sum_{k\in\mathcal{K}} w_k(p)\, \mathbf{F}(p+k), \tag{3}$$

where $\mathcal{K} = \{(0,0), (\pm 1, 0), (0, \pm 1)\}$. The MLP consists of two 1×1 convolutions separated by GELU and is applied independently at each pixel. The fused feature $\mathbf{F}$ is deliberately excluded from the predictor, so the direction is determined by event evidence rather than by the feature being realigned. A convolution maps the aggregated feature $\mathbf{F}'$, and the result is added residually to $\mathbf{F}$.

Softmax places $\mathbf{w}(p)$ on the probability simplex, making the aggregation at each pixel a convex combination of five fixed views. Unlike a conventional five-tap convolution, which applies one kernel across the image, COMPASS predicts a location-specific kernel from $[\mathbf{Y}, \mathbf{C}]$. Its five weight maps therefore encode a spatially varying choice among the centre and four axial directions. The fixed cross is deliberately more restrictive than a learned offset field. It covers the four elementary directions around each location, keeps every sample on the feature grid and makes the selected direction directly visible in the weight maps. The centre tap allows the module to preserve already aligned content, while the residual connection lets the decoder fall back to the original fused feature. Thus, COMPASS adds spatial adaptivity without requiring unconstrained coordinate regression.

3.4 IMPLEMENTATION

The three mechanisms are the only architectural differences from the controlled baseline we compare against (Fan et al., 2026): the backbone units, the stage widths $[64, 64, 128, 256]$, the loss (Charbonnier plus a frequency-domain term) and the optimiser schedule are shared, and the default configuration uses $(2, 2, 1, 1)$ blocks per stage while a larger configuration uses $(4, 2, 2, 2)$. Section 4 reports both, and every replacement ablation varies one mechanism while holding the other two and the training recipe fixed.

4 EXPERIMENTS

4.1 SETTINGS

Datasets and protocol. We evaluate on GoPro (Nah et al., 2017) with synthetic events, HighREV (Sun et al., 2023a) with real events, and NTURain (Chen et al., 2018) with synthetic rain. Deblurring uses six-channel event voxels, while deraining follows the released ten-channel EGVD protocol (Zhang et al., 2025). All outputs are evaluated with one unified script. The 37.10 dB TRM-UNet result in Table 1 comes from its official weights; the 37.18 dB baseline in Table 4 is retrained with our recipe and is used only for controlled attribution. Dataset splits, training hyperparameters and baseline protocols are provided in App. B.

4.2 COMPARISON WITH THE STATE OF THE ART

GoPro. Tab. 1 lists the event-guided methods on synthetic blur. Our model improves the best comparison result from 37.10 to 37.72 dB, and it is one of the two rows that beat AHDINet on all three metrics, with a $+0.74$ dB PSNR margin versus $+0.12$ dB for TRM-UNet, while the larger configuration of the same architecture reaches 37.97 dB. The gain is consistent across distortion, structure and perceptual quality: relative to TRM-UNet, our base model improves PSNR by 0.62 dB and SSIM by 0.0038, while reducing LPIPS from 0.02992 to 0.02570. Scaling the same design further adds 0.25 dB and lowers LPIPS to 0.02469. This monotonic improvement is important because it separates architectural scalability from a one-off gain at a single capacity.

Table 1: **Quantitative comparison on GoPro.** PSNR in dB. ◇ denotes a multi-frame method. The TRM-UNet row is obtained by evaluating its official pretrained weights with our unified script. **Bold with underline**, **bold** and underline mark the first, second and third best result of each column.

Model	PSNR↑	SSIM↑	LPIPS↓	Model	PSNR↑	SSIM↑	LPIPS↓
MAENet	34.50	0.9534	0.04291	AHDINet	36.98	0.9642	0.03058
EFNet	35.46	0.9561	0.03973	DuIntNet	37.00	0.9671	0.03102
EDMF	35.76	0.9596	0.03590	BDHNet	37.03	0.9618	0.03255
EIFNet	35.93	0.9561	0.04084	TRM-UNet	37.10	0.9645	0.02992
STCNet◇	36.35	0.9598	0.03648	Ours	**37.72**	**0.9683**	**0.02570**
MAT	36.67	0.9654	0.03029	Ours (large)	**37.97**	**0.9699**	**0.02469**

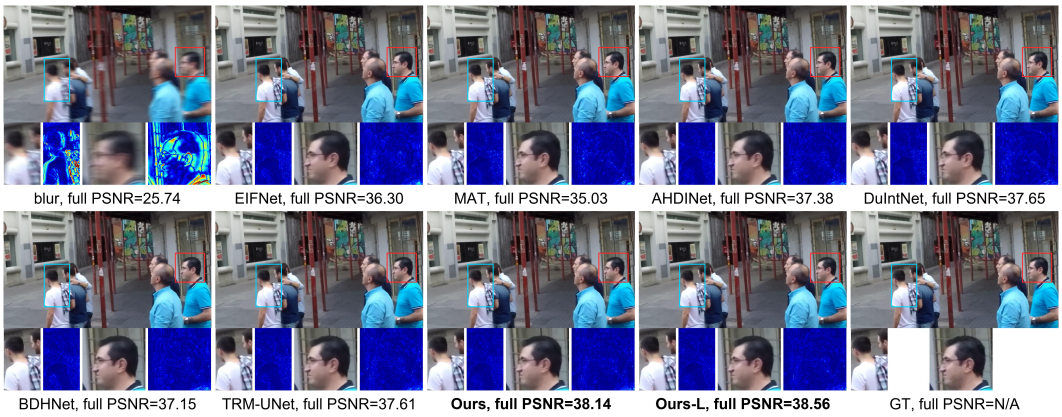

Figure 2: **Qualitative comparison on GoPro.** Each panel shows the full frame, a zoomed crop and the error map, with its full PSNR printed below.

Table 2: **Quantitative comparison on HighREV.** Rows marked *original* keep the loss of the original paper. **Bold** and underline mark the best and the second best result of each column.

Model	Loss	PSNR (dB)↑	SSIM↑	LPIPS↓
DuIntNet (Xiao et al., 2026)	original	35.30	0.9438	0.07398
BDHNet (Lin et al., 2025a)	original	35.76	0.9468	0.07195
EFNet (Sun et al., 2022)	original	36.40	0.9504	0.06450
AHDINet (Yang et al., 2025)	Charbonnier+FFT	35.11	0.9422	0.08523
AHDINet (Yang et al., 2025)	original	35.68	0.9467	0.07623
Ours	Charbonnier+FFT	**37.03**	**0.9528**	**0.06362**

HighREV. Tab. 2 reports the real-event setting, where the event stream is noisy and its reliability varies strongly between moving edges and flat regions. Among the quoted methods EFNet is the strongest at 36.40 dB, which our model improves by 0.63 dB; both AHDINet objectives are listed because the objective dominates its performance, and our model improves the better variant by 1.35 dB PSNR, 0.0061 SSIM and reduces LPIPS by 0.0126. These results show that the improvement persists in the real-event setting despite spatially varying event noise. The three metrics also agree on the ranking, even though the margin in LPIPS is smaller than that in PSNR. This consistency matters on HighREV because an improvement confined to PSNR could result from suppressing noisy events at the expense of perceptual detail. Instead, the simultaneous gains in PSNR, SSIM and LPIPS indicate that the model improves fidelity without trading away structural or perceptual quality.

NTURain. Tab. 3 reports deraining. The two event-guided video derainers that accept events together with a frame window, RainDancer-s and RainDancer, reach 37.19 and 38.19 dB, respectively. Our model reaches 46.38 dB, 0.90 dB above the strongest quoted competitor, and is also the best

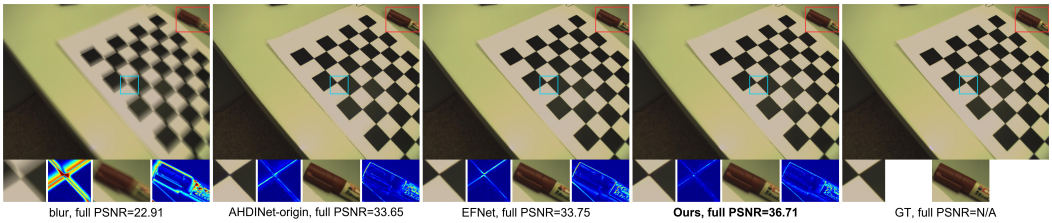

Figure 3: **Qualitative comparison on HighREV.** Each panel shows the full frame, a zoomed crop and the error map, with its full PSNR printed below.

Table 3: **Quantitative comparison on NTURain.** Deraining, PSNR in dB. $^{\diamond}$ marks a multi-frame input and $^{\circ}$ a method that takes no events; unmarked rows take one frame plus an event voxel. EGVD* resets its recurrent state for every input. **Bold with underline**, **bold** and underline mark the first, second and third best result of each column.

Model	PSNR$^{\uparrow}$	SSIM$^{\uparrow}$	LPIPS$_{\downarrow}$	Model	PSNR$^{\uparrow}$	SSIM$^{\uparrow}$	LPIPS$_{\downarrow}$
UVD$^{\diamond}$	31.34	0.9386	0.05416	VDMamba$^{\circ\diamond}$	39.79	0.9860	0.00610
RainDancer-s$^{\diamond}$	37.19	0.9685	0.01876	EGVD$^{*\diamond}$	44.06	0.9934	0.00241
DeLiVR$^{\circ\diamond}$	37.33	0.9755	0.01274	MPEVNet$^{\diamond}$	45.34	0.9944	0.00183
RainMamba$^{\circ\diamond}$	37.91	0.9739	0.01042	AHDINet	**45.48**	**0.9948**	**0.00162**
RainDancer$^{\diamond}$	38.19	0.9728	0.01425	Ours	**46.38**	**0.9954**	**0.00161**

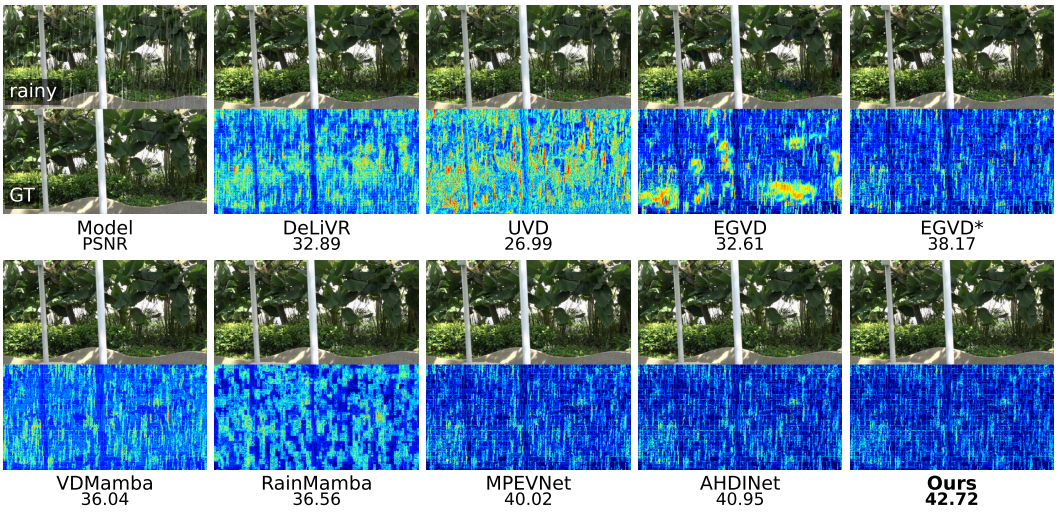

Figure 4: **Qualitative comparison on NTURain.** The leftmost column shows the rainy input and the ground truth; each remaining panel shows the full frame, a zoomed crop and the error map, with its patch PSNR printed below.

on SSIM and LPIPS. Two observations are worth recording: we report EGVD (Zhang et al., 2025) with its recurrent state reset for every input, since without the reset it collapses to a level that is not comparable; and LPIPS is saturated on this benchmark, so PSNR and SSIM carry the comparison. The NTURain result also tests whether the proposed interaction rule is specific to motion deblurring. Unlike GoPro and HighREV, rain produces repeated directional structures across the image, yet the same VIBE–COEX–COMPASS design improves over both single-frame AHDINet and multi-frame MPEVNet. The 0.90 dB gain over AHDINet is accompanied by a further SSIM increase from 0.9948 to 0.9954, suggesting that the improvement is not caused solely by a few high-error regions.

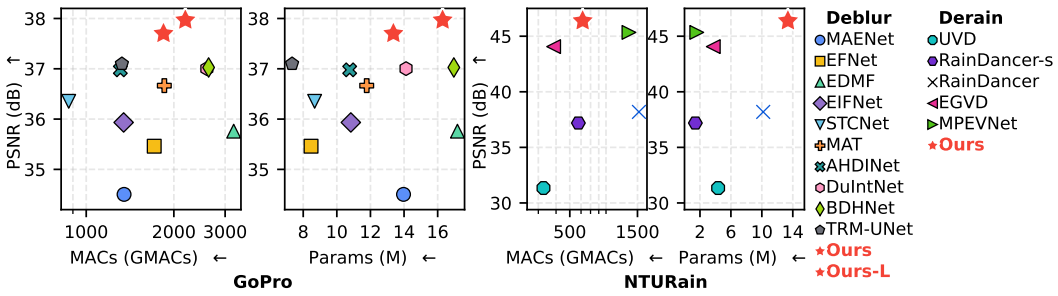

Figure 5: **Accuracy against cost.** Event-guided methods on GoPro (left pair) and NTURain (right pair); the vertical axis is PSNR and the horizontal axes are MACs (GMACs, log scale) and parameters (M), so the lower left is the cheapest. MACs are counted with `fvcore` (one multiply–accumulate counted as one) at each model's operating resolution (768×1280 on GoPro, 512×640 on NTURain, batch 1). Published PSNR values are quoted from the cited papers and ours are the models of Tabs. 1 and 3.

4.3 Efficiency

The three mechanisms are not free, and we quantify the cost rather than quote a single number. On GoPro the complete model uses 1842.6 GMACs and 13.37M parameters against 1326.0 GMACs and 7.34M for the retrained TRM-UNet, i.e. $1.4\times$ the computation and $1.8\times$ the parameters, for $+0.54$ dB over that partner ($+0.79$ dB in the larger configuration at 2189.9 GMACs and 16.29M). On NTURain, where the baseline is a recurrent or multi-frame design, the picture reverses: our single-frame model needs 615.7 GMACs against 1306.6 for MPEVNet and 1536.2 for RainDancer, while being 0.9–8.2 dB more accurate. Fig. 5 places all event-guided methods in the accuracy–cost plane: our models occupy the accuracy end of both benchmarks without leaving the cost range of the methods they are compared with. The two tasks reveal different efficiency regimes. On GoPro, event-native processing produces an accuracy–cost trade-off rather than a free improvement. On NTURain, however, the single-frame design uses less than half the MACs of MPEVNet and RainDancer while achieving higher PSNR.

4.4 Qualitative Results

GoPro. Figure 2 highlights the person's collar on the right, where competing methods leave softened or irregular boundaries. Our reconstruction preserves a cleaner, more continuous collar contour, and Ours (large) recovers the sharpest local structure with the lowest residual error.

HighREV. The real-event example in Fig. 3 contains both repeated high-contrast structure and small object boundaries. Our method reconstructs cleaner transitions at the black–white chessboard boundaries and produces a sharper, more complete edge around the marker cap. These details remain well defined despite the spatially varying noise of real events.

NTURain. In Fig. 4, our full-image error map is visibly cleaner than those of the competing methods, with substantially weaker residual responses across the scene. The enlarged crop also retains local image structure while suppressing directional rain streaks, consistent with the improvement in all three quantitative metrics.

4.5 Ablation Study

We ablate the three mechanisms on the full GoPro dataset (Nah et al., 2017). Starting from the controlled 37.18 dB baseline, VIBE yields the largest incremental gain, improving PSNR by 0.27 dB. Adding COEX raises performance from 37.45 to 37.59 dB, and COMPASS provides a further 0.13 dB to reach 37.72 dB. These are cumulative gains in this insertion order rather than independent estimates of each module's effect; together they improve the controlled baseline by 0.54 dB.

Table 4: **GoPro ablation.** The first block adds one mechanism at a time on top of the retrained controlled partner; the second replaces one component of the complete model. VIBE, COEX and COMPASS mark which of the three mechanisms is active, and all rows use the same recipe, seed and widths.

Setting	VIBE	COEX	COMPASS	PSNR (dB)$^{\uparrow}$
Baseline (hand mask and gate)	–	–	–	37.18
+ VIBE	✓	–	–	37.45
+ VIBE + COEX	✓	✓	–	37.59
Ours	✓	✓	✓	37.72
Replace one component of the complete model				
COEX without the projection	✓	✓	✓	37.57
VIBE with the activity map only	✓	✓	✓	37.52
COMPASS with global weights	✓	✓	✓	37.56

The replacement experiments test the defining choice within each mechanism while keeping the other two active. Removing the projection from COEX reduces PSNR by 0.15 dB, indicating that the gain does not come solely from adding another learned mixer. Replacing VIBE's four statistics with the activity map alone incurs the largest replacement drop, from 37.72 to 37.52 dB. This comparison isolates the value of polarity balance and temporal moments beyond total activity. Finally, replacing COMPASS's per-pixel weights with global weights lowers PSNR to 37.56 dB, supporting spatial adaptation over a single image-wide directional preference. The similar drops across the three replacements show that the complete result depends on the specific constraint or conditioning used by each mechanism, not merely on additional branches.

Cross-benchmark insight. Against the strongest comparison in each table, the base model gains 0.62 dB on synthetic-event GoPro, 0.63 dB on real-event HighREV and 0.90 dB on NTURain. The similar GoPro and HighREV margins suggest that the improvement is not tied to simulated events, while the NTURain result extends the same interaction rule from motion blur to directional rain. Together with the ablations, this pattern supports a division of labour in which VIBE preserves event structure, COEX controls cross-modal contribution and COMPASS resolves local directional ambiguity. The secondary metrics refine this interpretation. On GoPro and HighREV, PSNR gains are accompanied by higher SSIM and lower LPIPS, indicating that the improvement is not obtained by optimising pixel fidelity alone. On NTURain, LPIPS values are already close to saturation, yet the method still improves PSNR and SSIM over AHDINet. The large configuration further improves every GoPro metric, providing evidence that the design continues to benefit from additional capacity.

5 Conclusion

We presented three complementary mechanisms for event-guided restoration. VIBE replaces a hand-designed activity mask with polarity-aware event processing conditioned on measured temporal statistics. COEX replaces independent gates with a per-channel mixer projected toward the doubly stochastic manifold, encouraging approximately sum-preserving exchange at every fusion site. COMPASS uses event-conditioned weights over five fixed shifts to introduce directional readout without offset regression. Together, these mechanisms improve the controlled GoPro baseline by 0.54 dB and its larger configuration by 0.79 dB, while achieving 37.03 dB on HighREV and 46.38 dB on NTURain.

The method also has clear limitations. The finite COEX projection is only approximately doubly stochastic and does not model a cost for transferring information between channels. VIBE increases the cost of the event branch, and the overall formulation assumes degradations that are integrated during exposure and reflected in the event stream. We therefore restrict our claims to motion blur and rain; extending the exchange to cost-aware transport and broader degradation models remains future work.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for drafting and revising parts of the manuscript, for writing analysis and plotting scripts, for surveying related work through the material collected in our repository, and for discussing the design of the ablation matrix. We have not used generative AI tools for generating experimental data or results, for data collection or annotation, or for producing figures and tables: every PSNR, SSIM and LPIPS value reported here comes from real training and evaluation runs, all datasets are public benchmarks processed by our own pipelines, and all figures are produced by our own scripts and slides. We have reviewed all AI-assisted work: the analysis scripts were executed and their outputs verified against the released event voxel files, every number in the tables was checked against the evaluation logs, and the manuscript text was checked against the cited papers. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

All models are trained and evaluated with the recipe, patch size, optimiser and schedule described in Sec. 4.1; the architecture and implementation details (widths, block configuration, the seven fusion sites, the Sinkhorn projection and the five-tap readout) are given in Sec. 3.4. The controlled comparisons of Sec. 4.2 and the ablation rows of Sec. 4.5 share data, patches, optimiser and schedule with the models reported in the main tables, so each reported gain is attributable to the component that was changed. Part of the code (configurations, model definitions and evaluation scripts) is provided as supplementary material; the complete code and the trained checkpoints will be released upon acceptance. Training the largest configuration took about three days on four A100 or four RTX 6000 Ada GPUs on HighREV, and less time on the other benchmarks; the per-sample qualitative comparisons are in App. A.

REFERENCES

Jie Chen, Cheen-Hau Tan, Junhui Hou, Lap-Pui Chau, and He Li. Robust Video Content Alignment and Compensation for Rain Removal in a CNN Framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6286–6295, 2018.

Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Revitalizing Convolutional Network for Image Restoration. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(12): 9423–9438, 2024.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 764–773, 2017. doi: 10.1109/ICCV.2017.89.

D. Fan, X. Tang, Q. Chen, and F. Xu. Trm-unet: An efficient event-guided motion deblurring network. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 9047–9051, 2026. doi: 10.1109/ICASSP55912.2026.11462796.

Guangsha Guo, Shilong Jing, Yuchen Zhao, Hengyi Lv, Yisa Zhang, and Yang Feng. Event-driven multimodal fusion for image motion deblurring. *Expert Systems with Applications*, 296:129044, 2026.

Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. MambaIRv2: Attentive State Space Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28124–28133, 2025.

Kui Jiang, Runzhe Li, Zhaocheng Yu, Guanglu Sun, Junjun Jiang, and Xianming Liu. Raindancer: Rgb-event video deraining with rain-oriented spiking dynamics. *arXiv preprint arXiv:2607.13802*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Boyun Li, Haiyu Zhao, Wenxin Wang, Peng Hu, Yuanbiao Gou, and Xi Peng. Mair: A locality- and continuity-preserving mamba for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7491–7501, 2025.

Xiaopeng Lin, Yulong Huang, Hongwei Ren, Zunchang Liu, Hongxiang Huang, Yue Zhou, Haotian Fu, and Bojun Cheng. Clearsight: Human vision-inspired solutions for event-based motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7462–7471, 2025a.

Yu-Cheng Lin, Yu-Syuan Xu, Hao-Wei Chen, Hsien-Kai Kuo, and Chun-Yi Lee. Eamamba: Efficient all-around vision state space model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11708–11719, 2025b.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 257–265, 2017.

Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Proceedings of the Conference on Robot Learning (CoRL)*, volume 87, pp. 969–982, 2018.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 412–428, 2022.

Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18043–18052, 2023a.

Lei Sun, Andrea Alfarano, Peiqi Duan, Shaolin Su, Kaiwei Wang, Boxin Shi, Radu Timofte, Danda Pani Paudel, Luc Van Gool, Qinglin Liu, Wei Yu, Xiaoqian Lv, Lu Yang, Shuigen Wang, Shengping Zhang, Xiangyang Ji, Long Bao, Yuqiang Yang, Jinao Song, Ziyi Wang, Shuang Wen, Heng Sun, Kean Liu, Mingchen Zhong, Senyan Xu, Zhijing Sun, Jiaying Zhu, Chengjie Ge, Xingbo Wang, Yidi Liu, Xin Lu, Xueyang Fu, Zheng-Jun Zha, Dawei Fan, Dafeng Zhang, Yong Yang, Siru Zhang, Qinghua Yang, Hao Kang, Huiyuan Fu, Heng Zhang, Hongyuan Yu, Zhijuan Huang, Shuoyan Wei, Feng Li, Runmin Cong, Weiqi Luo, Mingyun Lin, Chenxu Jiang, Hongyi Liu, Lei Yu, Weilun Li, Jiajun Zhai, Tingting Lin, Shuang Ma, Sai Zhou, Zhanwen Liu, Yang Wang, Eiffel Chong, Nuwan Bandara, Thivya Kandappu, Archan Misra, Yihang Chen, Zhan Li, Weijun Yuan, Wenzhuo Wang, Boyang Yao, Zhanglu Chen, Yijing Sun, Tianjiao Wan, Zijian Gao, Qisheng Xu, Kele Xu, Yukun Zhang, Yu He, Xiaoyan Xie, Tao Fu, Yashu Gautamkumar Patel, Vihar Ramesh Jain, Divesh Basina, Rishik Ashili, Manish Kumar Manjhi, Sourav Kumar, Prinon Benny, Himanshu Ghunawat, B Sri Sairam Gautam, Anett Varghese, and Abhishek Yadav. NTIRE 2025 Challenge on Event-Based Image Deblurring: Methods and Results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1315–1332, 2025a.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Kaihao Zhang, Meiyu Liang, and Xiaochun Cao. Event-aware video deraining via multi-patch progressive learning. *IEEE Transactions on Image Processing*, 32:3040–3053, 2023b.

Shangquan Sun, Wenqi Ren, Juxiang Zhou, Shu Wang, Jianhou Gan, and Xiaochun Cao. Semi-supervised state-space model with dynamic stacking filter for real-world video deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26114–26124, 2025b.

Shuning Sun, Jialang Lu, Xiang Chen, Jichao Wang, Dianjie Lu, Guijuan Zhang, Guangwei Gao, and Zhuoran Zheng. DeliVR: Differential spatiotemporal lie bias for efficient video deraining. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Zhijing Sun, Xueyang Fu, Longzhuo Huang, Aiping Liu, and Zheng-Jun Zha. Motion aware event representation-driven image deblurring. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 418–435, 2024.

Jin Wang, Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Unsupervised Video Deraining with An Event Camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10831–10840, 2023.

Hongtao Wu, Yijun Yang, Huihui Xu, Weiming Wang, Jinni Zhou, and Lei Zhu. Rainmamba: Enhanced locality learning with state space models for video deraining. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pp. 7881–7890, 2024.

Zeyu Xiao, Zhuoyuan Li, Yang Zhao, Yu Liu, Zhao Zhang, and Wei Jia. Learning dual modality interactions for event-based motion deblurring. *IEEE Transactions on Multimedia*, 28:3652–3666, 2026.

Senyan Xu, Zhijing Sun, Mingchen Zhong, Chengzhi Cao, Yidi Liu, Xueyang Fu, and Yan Chen. Motion-adaptive transformer for event-based image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 8942–8950, 2025.

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 ACM International Conference on Multimedia (ACM MM)*, pp. 8327–8335, 2023.

Wen Yang, Jinjian Wu, Jupo Ma, Leida Li, and Guangming Shi. Motion deblurring via spatial-temporal collaboration of frames and events. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 6531–6539, 2024.

Wen Yang, Jinjian Wu, Leida Li, Weisheng Dong, and Guangming Shi. Asymmetric hierarchical difference-aware interaction network for event-guided motion deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 9265–9273, 2025.

Yueyi Zhang, Jin Wang, Wenming Weng, Xiaoyan Sun, and Zhiwei Xiong. Egvd: Event-guided video deraining. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):14092–14106, 2025.

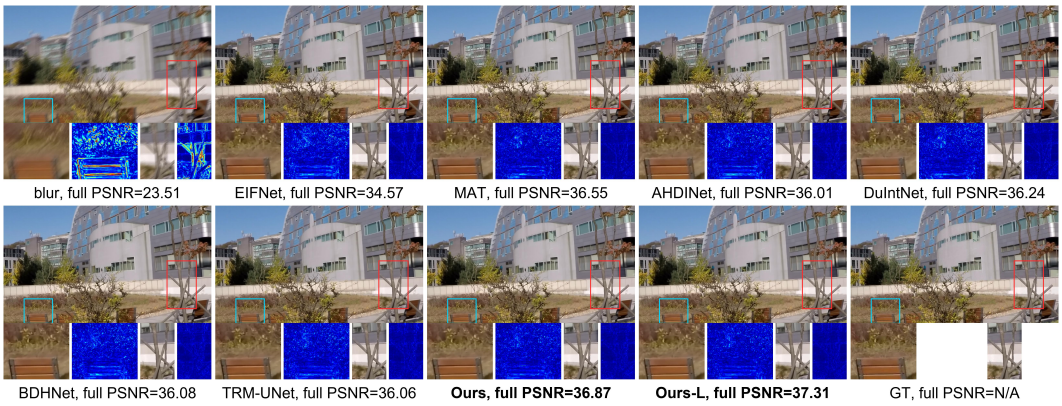

Figure 6: **Qualitative comparison on GoPro, second sequence.** Same panel layout as Fig. 2.

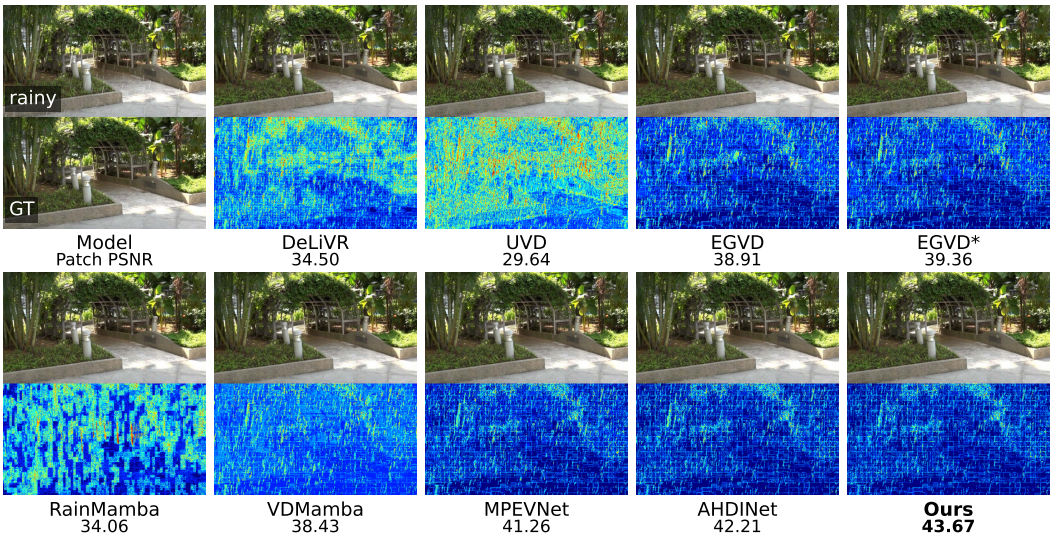

Figure 7: **Qualitative comparison on NTURain, another sequence.** Same panel layout as Fig. 4.

A Additional qualitative results

Fig. 6 repeats the comparison of Fig. 2 on GoPro, and Fig. 7 does the same for Fig. 4 on NTURain.

B Experimental details

Datasets. GoPro (Nah et al., 2017) contains 22 training and 11 test clips. We generate its events with ESIM (Rebecq et al., 2018) following Yang et al. (2023), and discard the first and last frames of each test sequence during evaluation. HighREV (Sun et al., 2023a) contains 20 training, 5 validation and 4 test clips. Because sharp ground truth is unavailable for its test split, we report the validation split. NTURain (Chen et al., 2018) contains 8 training and 8 test clips; when comparing with seven-frame methods, we exclude the first and last three frames. We retrain UVD, EGVD, MPEVNet, RainDancer and RainDancer-s using their official implementations and the released ten-channel EGVD event data (Zhang et al., 2025), because pretrained checkpoints are unavailable for most of these methods.

Training. All models are trained for 300k iterations with Adam (Kingma & Ba, 2015), a batch size of 4, and an initial learning rate of 2×10^{-4} cosine-annealed to 10^{-6}. We apply random horizontal flips and rotations, each with probability 0.5. The input patch size is 512×512 for deblurring and

384×384 for deraining. Checkpoints are evaluated every 1k iterations, and we select the model with the highest validation PSNR.

Comparison protocol. The state-of-the-art tables include event-guided deblurring methods (Sun et al., 2024; 2022; Guo et al., 2026; Yang et al., 2023; 2024; Xu et al., 2025; Yang et al., 2025; Xiao et al., 2026; Sun et al., 2025a; Fan et al., 2026) and event-guided derainers (Wang et al., 2023; Sun et al., 2023b; Zhang et al., 2025; Sun et al., 2025b; Wu et al., 2024; Sun et al., 2026; Jiang et al., 2026). The GoPro state-of-the-art comparison evaluates the official TRM-UNet weights with our unified script and obtains 37.10 dB. For the controlled ablation, we retrain TRM-UNet with the same data, patches, optimiser, schedule and seed as our model, obtaining 37.18 dB. All ablation gains are measured against this retrained baseline.