MAMBAFOURIER: A PHYSICS-INSPIRED SPECTRAL FUSION MODEL FOR LOW-LIGHT EVENT–RGB SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-light perception remains a fundamental bottleneck, as insufficient brightness causes texture to collapse and appearance-based segmentation to fail catastrophically. Event cameras remain informative in these regimes, yet existing RGB–Event fusion modules are generic spatial or token mixers that ignore the modalities' spectral asymmetry: events are sparse high-frequency temporal derivatives of log intensity, whereas frames carry reliable low-frequency structure but noise-dominated high frequencies under low light. We introduce **MambaFourier**, a dual-branch state-space segmentation network whose fusion operator is physics-derived. At each scale, a **Wave–Fourier Transform Fusion (WFTF)** block treats the two feature maps as coupled damped wave fields. Since the Laplacian is diagonal in the Fourier basis, the damped wave equation has a closed-form per-wavenumber solution: WFTF performs one real FFT, applies a learnable per-channel wave propagator, a band-limited free-form filter, and a reciprocal 2×2 cross-modal coupling, then one inverse FFT. Spectral parameters are indexed by mode, enabling resolution transfer, and WFTF is an exact identity at initialisation. MambaFourier reaches 79.39% mIoU on DDD17 ($+1.83\%$ over MambaSeg), 75.65% on DSEC-Semantic ($+4.61\%$ over ESC), 70.40% on DERS-XS ($+3.30\%$ over ESC), and 56.51% on DSEC-Xtrm ($+5.64\%$ over the previous best). Under synthetic rain and fog on DSEC, it attains 73.7% mIoU versus 72.5% for MambaSeg.

1 INTRODUCTION

Semantic segmentation is now a mature capability under benign capture conditions, with transformer and convolutional architectures reporting strong results on standard driving benchmarks (Xie et al., 2021; Guo et al., 2022; Thisanke et al., 2023). That maturity is deceptive. The assumptions behind an RGB encoder – adequate exposure, limited motion blur, a usable signal-to-noise ratio – are precisely the assumptions that fail in the deployment scenarios that matter most: night-time driving, tunnel exits, sudden headlight glare, aggressive ego-motion. In these regimes the frame does not merely become noisy; whole regions lose the texture gradients on which appearance-based features depend, and segmentation collapses in a way that no amount of decoder capacity can repair.

Event cameras offer a genuinely complementary measurement channel. Each pixel fires asynchronously on logarithmic brightness change (Lichtsteiner et al., 2008; Gallego et al., 2022), which yields microsecond latency, a dynamic range beyond $120\,\mathrm{dB}$, and near-immunity to motion blur (Gehrig & Scaramuzza, 2024). Crucially, the failure modes are disjoint: an event stream is richest exactly where a frame is poorest, because both are driven by scene motion and contrast rather than by absolute illumination. This complementarity has motivated a substantial line of event-assisted and event-only segmentation work (Alonso & Murillo, 2019; Sun et al., 2022; Zhao et al., 2025; Kong et al., 2024), and a parallel line of RGB-Event fusion architectures (Zhang et al., 2023a;b; Xie et al., 2024a; Gu et al., 2026).

Existing fusion modules cross-attention, token substitution, channel/spatial recalibration are generic mixers that leave the burden of distinguishing the two streams entirely to the data. The distinction is best stated in the frequency domain: an event voxel grid concentrates energy at high spatial

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.

wavenumbers (contours, poles, lane markings) and carries almost nothing at low ones, since static regions emit no events, while an RGB frame carries reliable low-wavenumber structure that persists even under low light, whose high-frequency band is instead overwhelmed by shot noise. A fusion operator that mixes *spatially* must rediscover this band structure through many layers; one that mixes *spectrally* can express it directly. We therefore ask what the two feature maps would do as physical fields allowed to interact, modelling each modality's stage feature as a damped scalar wave field $u(\boldsymbol{x}, t)$: wave propagation spreads reliable evidence outward over a frequency-dependent extent while damping suppresses the noisy high-wavenumber band, and the damped wave equation is diagonal in the Fourier basis, so each wavenumber evolves independently with a closed-form solution – no PDE integration, no time-step unrolling, no neural-ODE-style stability constraints. We evaluate the analytic propagator in one shot, at the cost of a pair of FFTs.

This yields the **Wave–Fourier Transform Fusion (WFTF)** block, the core of **MambaFourier**. At each pyramid scale, WFTF projects the image and event features into a displacement field and a velocity field, transforms them with a single batched real FFT, and applies three per-frequency operations: (i) a *spectral wave propagator* whose speed, damping and propagation time are learned per channel, so different channels specialise to different frequency bands; (ii) a *band-limited learnable Fourier filter* in the spirit of Fourier neural operators and global filter networks (Li et al., 2021; Rao et al., 2021; Guibas et al., 2022), which adds free-form capacity on the lowest modes while letting out-of-band coefficients pass through unchanged; and (iii) a *reciprocal cross-modal coupling* that exchanges energy between the two fields mode by mode, in the manner of two coupled oscillators. A single inverse FFT returns to the spatial domain.

The encoder is a pair of GroupMamba backbones (Shaker et al., 2025). We choose a state-space backbone deliberately: selective SSMs (Gu & Dao, 2024; Liu et al., 2024; Zhu et al., 2024) give global receptive field at linear cost, which pairs naturally with a global spectral fusion operator, and their linear scaling keeps the dual-branch design affordable at the high input resolutions that event data reward.

Contributions

1. **A spectral fusion operator built on a closed-form PDE solution.** Wave–Fourier Fusion models paired RGB and event features as damped wave fields and propagates them with the exact solution of the damped wave equation in the Fourier basis. The propagator is parameterised by three interpretable per-channel quantities (wave speed, damping and propagation time). The block's filter and coupling weights are indexed by mode number, so one configuration applies at every feature-map size, and the block is an exact identity at initialisation, so it can be inserted into a pretrained dual-branch encoder without a warm-up schedule.
2. **Per-frequency reciprocal cross-modal coupling.** We formulate cross-modal exchange as a symmetric 2×2 complex matrix applied independently at each retained wavenumber, so events can donate high-frequency structure to RGB bands destroyed by low light without contaminating the low-frequency bands where RGB remains trustworthy.
3. **State-of-the-art results, with the largest margins where conditions are worst.** Combining dual GroupMamba encoders with stage-wise WFTF, MambaFourier attains 79.39% mIoU on DDD17 and 75.65% mIoU on DSEC-Semantic, and improves the previous best by 5.64% mIoU on the degraded DSEC-Xtrm benchmark and 3.30% mIoU on the true-labelled synthetic DERS-XS benchmark – evidence that the gain comes from resilience rather than from benchmark-specific tuning.

2 Related Work

2.1 Event-based and event–RGB semantic segmentation

Event-based segmentation was opened up by Ev-SegNet (Alonso & Murillo, 2019), which rendered events into a six-channel image-like tensor and applied a conventional CNN on DDD17 (Binas et al., 2017). Because dense per-pixel labels for event streams are expensive, much of the subsequent literature is about supervision rather than architecture: ESS transfers knowledge from labelled images to unlabelled events via unsupervised domain adaptation (Sun et al., 2022), OpenESS pursues open-vocabulary event scene understanding (Kong et al., 2024), and CMESS aligns modalities

with attention-based soft alignment (Xie et al., 2024b). Efficiency-oriented work replaces the ANN backbone with spiking encoders (Zhang et al., 2025; Long et al., 2025) or hybrid ANN–SNN designs (Li et al., 2025). A separate thread exploits the structural affinity between events and object boundaries: ESEG uses explicit edge-semantic guidance for uni-modal event segmentation (Zhao et al., 2025), and ESC (Bao et al., 2025) builds a VQ-VAE edge dictionary as a shared discrete latent space between the modalities, deriving per-pixel uncertainties from the re-coded edge distribution.

For paired event–RGB input, CMX (Zhang et al., 2023a) and CMNeXt (Zhang et al., 2023b) established the dominant template: a transformer encoder per modality plus cross-modal rectification and fusion at every stage. TokenFusion (Wang et al., 2022) instead prunes uninformative tokens and substitutes them with projections from the other stream. EISNet (Xie et al., 2024a) recalibrates and aggregates the two streams at multiple scales; SAM-Event-Adapter (Yao et al., 2024) adapts a segmentation foundation model; OmniSegmentor (Yin et al., 2025) generalises to arbitrary modality sets; and BRENet (Yao et al., 2025) tackles the spatial misalignment between the two sensors with flow-guided registration. The most closely related recent system is MambaSeg (Gu et al., 2026), which is, to our knowledge, the first to use a state-space backbone for image–event segmentation.

2.2 State space models for vision

Structured state space models (Gu et al., 2022) and their selective variant Mamba (Gu & Dao, 2024) model long sequences in linear time while retaining a global receptive field, which makes them an appealing alternative to quadratic self-attention for dense prediction. Vision Mamba (Zhu et al., 2024) introduced a bidirectional SSM backbone, and VMamba (Liu et al., 2024) proposed a 2D selective scan to resolve the direction-sensitivity of flattening an image into a 1D sequence. GroupMamba (Shaker et al., 2025) splits channels into groups scanned in four complementary directions and stabilises training with a distilled objective, giving a hierarchical four-stage backbone with a favourable accuracy/throughput trade-off. We adopt GroupMamba as our per-modality encoder.

2.3 Spectral and Fourier neural architectures

Operating in the frequency domain has a long history in neural Partial Differential Equation (PDE) solvers and efficient token mixing. The Fourier Neural Operator (Li et al., 2021) parameterises an integral kernel as a band-limited multiplication in Fourier space, which makes it resolution-independent – the property we exploit for cross-dataset transfer. GFNet (Rao et al., 2021) replaces self-attention with a learnable global filter on the 2D spectrum, AFNO (Guibas et al., 2022) adds block-diagonal adaptivity and soft-thresholding, and FNet (Lee-Thorp et al., 2022) shows that even an unparameterised DFT is a competitive token mixer. These models learn free-form frequency responses. Our propagator instead derives its response from the analytic solution of a PDE, so the learned degrees of freedom are three physically interpretable scalars per channel rather than a dense complex tensor.

3 Method

3.1 Overview

Let $\mathcal{I} \in \mathbb{R}^{3\times H\times W}$ denote an RGB frame and $\mathcal{E} \in \mathbb{R}^{T\times H\times W}$ its corresponding event voxel grid with T temporal bins (Zhu et al., 2018). We use separate GroupMamba-Small encoders (Shaker et al., 2025) to extract four-stage features. The encoders share an architecture but have independent weights; their directional selective scans capture long-range context within each modality. We initialise the event stem by replicating the three-channel pretrained weights and scaling them by $3/T$. At each stage, WFTF refines both feature maps, which then contribute to the decoder and the subsequent encoder stages:

$$(\tilde{X}_i^{\mathcal{I}}, \tilde{X}_i^{\mathcal{E}}) = \mathrm{WFF}_i(X_i^{\mathcal{I}}, X_i^{\mathcal{E}}), \quad F_i = \mathrm{Conv}_{1\times 1}([\tilde{X}_i^{\mathcal{I}} \,\|\, \tilde{X}_i^{\mathcal{E}}]), \quad X_{i+1}^m = \mathrm{Down}_i(X_i^m + \tilde{X}_i^m). \tag{1}$$

The fused map $F_i \in \mathbb{R}^{C_i\times H_i\times W_i}$ is sent to the decoder, while the residual feedback into Down_i, the patch-merging layer, lets cross-modal evidence influence all deeper stages of both encoders (Figure 1). The remainder of this section derives the WFTF block: the wave model (Section 3.2), its three spectral operators (Sections 3.3–3.5), and their assembly (Section 3.6).

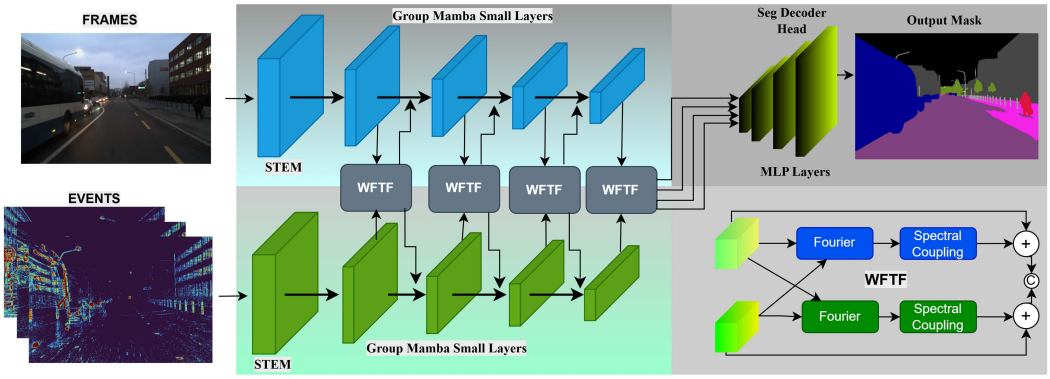


Figure 1: Overview of MambaFourier. Parallel GroupMamba branches encode image and event inputs. At four scales, the Wave–Fourier Fusion (WFTF) block treats the two feature maps as coupled damped wave fields and computes their interaction in closed form in the spectral domain.

3.2 Coupled damped wave fields

We treat each channel of a stage feature map as a scalar field on the 2D grid, with displacement $u_0^m(\boldsymbol{x})$ and velocity $v_0^m(\boldsymbol{x})$ predicted from the features (Section 3.6), evolving over a horizon $\tau > 0$ under

$$\partial_\tau^2 u \;-\; 2\gamma\,\nabla^2 \partial_\tau u \;+\; \gamma^2 \nabla^4 u \;=\; c^2\,\nabla^2 u, \qquad u(\cdot,0) = u_0, \quad \partial_\tau u(\cdot,0) = v_0, \tag{2}$$

with wave speed $c > 0$ and damping $\gamma > 0$. Strain-rate (Kelvin–Voigt) damping $-2\gamma\nabla^2\partial_\tau u$ attenuates high wavenumbers faster than low ones, suppressing the shot-noise-dominated band of low-light frames; the biharmonic term keeps every mode underdamped. The Fourier transform diagonalises the Laplacian, so each wavenumber $\boldsymbol{k}$ evolves as an independent oscillator,

$$\ddot{\hat{u}}_{\boldsymbol{k}} + 2\alpha_{\boldsymbol{k}}\,\dot{\hat{u}}_{\boldsymbol{k}} + \big(\omega_{\boldsymbol{k}}^2 + \alpha_{\boldsymbol{k}}^2\big)\hat{u}_{\boldsymbol{k}} = 0, \qquad \alpha_{\boldsymbol{k}} = \gamma\|\boldsymbol{k}\|^2, \quad \omega_{\boldsymbol{k}} = c\|\boldsymbol{k}\|, \tag{3}$$

with roots $-\alpha_{\boldsymbol{k}} \pm i\,\omega_{\boldsymbol{k}}$ and exact solution

$$\begin{aligned} \hat{u}(\boldsymbol{k},\tau) &= A(\boldsymbol{k})\,\hat{u}_0(\boldsymbol{k}) + B(\boldsymbol{k})\big(\hat{v}_0(\boldsymbol{k}) + \alpha_{\boldsymbol{k}}\hat{u}_0(\boldsymbol{k})\big), \\ A(\boldsymbol{k}) &= e^{-\alpha_{\boldsymbol{k}}\tau}\cos(\omega_{\boldsymbol{k}}\tau), \qquad B(\boldsymbol{k}) = e^{-\alpha_{\boldsymbol{k}}\tau}\,\frac{\sin(\omega_{\boldsymbol{k}}\tau)}{\omega_{\boldsymbol{k}}}. \end{aligned} \tag{4}$$

A single evaluation of equation (4) *is* the propagation: the envelope suppresses high wavenumbers, and the phase spreads evidence over an extent set by $c\tau$. We compute $\sin(\omega\tau)/\omega$ as $\tau\,\mathrm{sinc}(\omega\tau/\pi)$, which is stable at $\boldsymbol{k} = \boldsymbol{0}$.

3.3 Spectral wave propagator

Each channel of each modality owns a triple kept positive by a softplus ζ, and the wavenumber grid uses normalised frequencies:

$$\begin{aligned} &c = \zeta(c_{\mathrm{raw}}), \quad \tau = \zeta(\tau_{\mathrm{raw}}), \quad \gamma = \zeta(\gamma_{\mathrm{raw}}), \\ &\boldsymbol{k} = 2\pi\,(f_y, f_x), \quad f_y \in \mathrm{fftfreq}(H_i), \quad f_x \in \mathrm{rfftfreq}(W_i), \end{aligned} \tag{5}$$

so $\|\boldsymbol{k}\| \in [0, \sqrt{2}\,\pi]$ at every resolution, initialised at $c \approx \tau \approx 1.0$, $\gamma \approx 0.13$. Given displacement and velocity spectra $\hat{U}^m, \hat{V}^m$, the propagated spectrum is

$$\hat{P}^m = A^m\,\hat{U}^m + B^m\big(\hat{V}^m + \gamma^m \hat{U}^m\big). \tag{6}$$

Using γ instead of $\alpha_{\boldsymbol{k}}$ in the offset leaves A and B unchanged, so equation (6) remains the exact solution of equation (3) for the effective initial velocity $\hat{V}^m + (\gamma - \alpha_{\boldsymbol{k}})\hat{U}^m$; since $\hat{V}^m$ is learned, this only reparameterises it. Channels specialise freely: small γ and large τ give a broadband long-range propagator, large γ a low-pass smoother (Figure 2b), making the frequency decoupling explicit and inspectable after training.

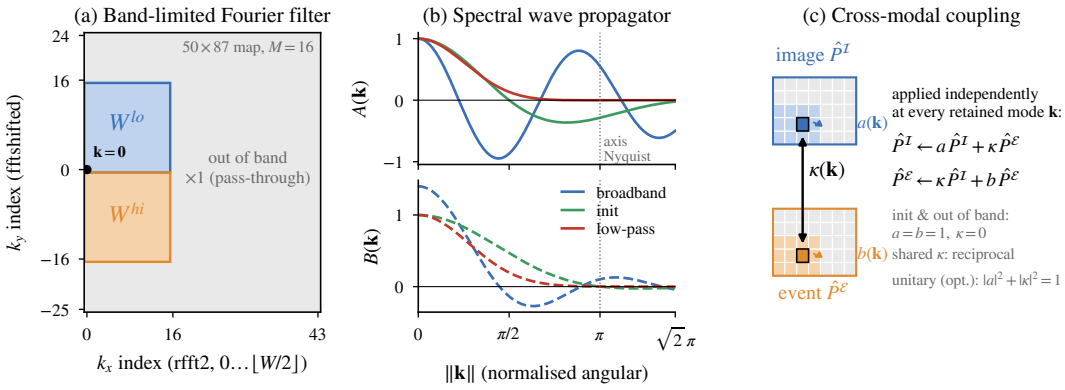


Figure 2: (a) Band-limited learnable Fourier filter on the rfft2 grid. (b) Spectral wave propagator responses $A(\boldsymbol{k})$, $B(\boldsymbol{k})$ for three learned parameter settings. (c) Reciprocal cross-modal coupling via a shared $\kappa(\boldsymbol{k})$.

3.4 Band-limited learnable Fourier filter

Following FNO and GFNet (Li et al., 2021; Rao et al., 2021), we add free-form capacity with a per-channel complex multiplier on the lowest M modes per axis, using two blocks $W^{m,lo}, W^{m,hi} \in \mathbb{C}^{C\times M\times M}$ for the non-negative and negative row frequencies:

$$\Phi^m(k_y,k_x)=\begin{cases} W^{m,lo}[k_y,k_x], & 0\le k_y<m_h,\ \ 0\le k_x<m_w,\\ W^{m,hi}[k_y,k_x], & -m_h\le k_y<0,\ \ 0\le k_x<m_w,\\ 1, & \text{otherwise,}\end{cases} \tag{7}$$

with $m_h=\min(M,\lfloor H_i/2\rfloor)$, $m_w=\min(M,\lfloor W_i/2\rfloor+1)$, applied as $\hat{P}^m\leftarrow\Phi^m\odot\hat{P}^m$. Out-of-band coefficients pass through unchanged: zeroing them yields a hard low-pass filter that damages thin classes such as poles, sign posts and lane markings. Because weights are indexed by mode number rather than grid position, the same parameters apply to a 50×87 DDD17 map and a 110×160 DSEC map without interpolation. We use $M=16$ at every stage.

3.5 Cross-modal spectral coupling

Two wave fields that share a medium exchange energy. We model this exchange as a reciprocal 2×2 complex matrix applied independently at each retained mode,

$$\begin{bmatrix}\hat{P}^{\mathcal{I}}(\boldsymbol{k})\\ \hat{P}^{\mathcal{E}}(\boldsymbol{k})\end{bmatrix}\leftarrow\begin{bmatrix}a(\boldsymbol{k}) & \kappa(\boldsymbol{k})\\ \kappa(\boldsymbol{k}) & b(\boldsymbol{k})\end{bmatrix}\begin{bmatrix}\hat{P}^{\mathcal{I}}(\boldsymbol{k})\\ \hat{P}^{\mathcal{E}}(\boldsymbol{k})\end{bmatrix},\qquad a,b,\kappa\in\mathbb{C}^{C\times M\times M}, \tag{8}$$

where a, b and κ use the same two-block band layout as equation (7), with out-of-band values $a=b=1$ and $\kappa=0$. The shared off-diagonal entry encodes reciprocity: the exchange is a property of the mode, not a directed transfer from one modality to the other. Because κ is indexed by wavenumber, the network can open the channel at high $\|\boldsymbol{k}\|$, where events are informative and a low-light frame is not, while keeping it nearly closed at low $\|\boldsymbol{k}\|$, where the frame still carries region identity that events cannot supply. An optional normalisation divides a, b and κ by $\sqrt{|a|^2+|\kappa|^2}$ per mode, which bounds the gain of the exchange; it is disabled by default.

3.6 The WFTF block

Given the stage features $X^{\mathcal{I}}, X^{\mathcal{E}}\in\mathbb{R}^{C\times H\times W}$ (stage index omitted), WFTF first produces the initial conditions of the two fields,

$$\begin{aligned}\left[u_0^m;\,v_0^m\right]&=\mathrm{Proj}^m_{1\times1}\Big(\mathrm{DWConv}^m_{3\times3}\big(\mathrm{LN}_{2d}(X^m)\big)\Big),\\ \hat{U}^m=\mathcal{F}\big(u_0^m\big),&\qquad \hat{V}^m=\mathcal{F}\big(v_0^m\big),\end{aligned} \tag{9}$$

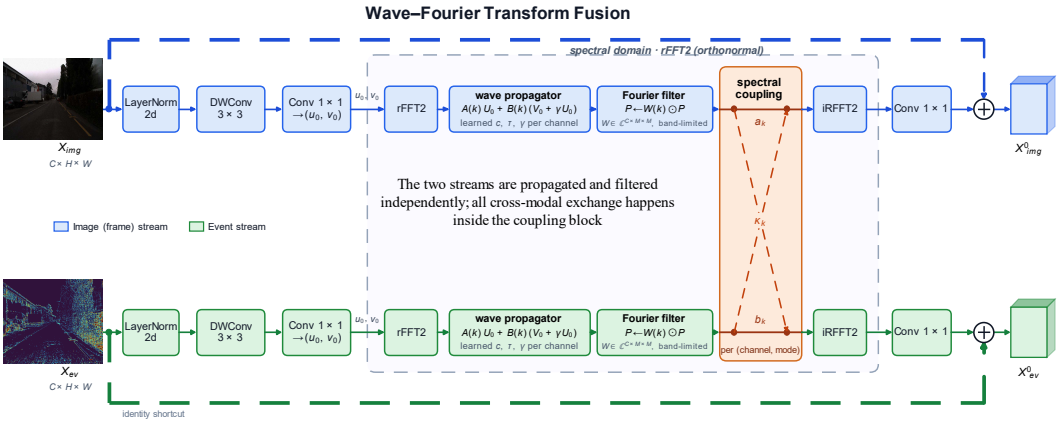


Figure 3: The Wave–Fourier Fusion block. Each stream yields displacement u_0 and separate velocity v_0; one batched real FFT performs propagation, filtering, and coupling, and one inverse FFT returns to the spatial domain. Zero-initialised LayerScale makes the block an identity at step 0.

where $\mathcal{F}$ is the orthonormal rfft2. It then propagates each spectrum with equation (6), filters it with equation (7), couples the two with equation (8), and returns to the spatial domain through a zero-initialised LayerScale residual,

$$\tilde{X}^m = X^m + \lambda^m \odot \mathrm{Conv}^m_{1\times1}\Big(\mathcal{F}^{-1}\big(\hat{P}^m\big)\Big), \qquad \lambda^m \in \mathbb{R}^C, \;\; \lambda^m = \mathbf{0} \text{ at initialisation.} \tag{10}$$

Displacement and the learned velocity feature come from separate output channels of the projection rather than from one shared map. The latter gives the propagator a second, independent initial condition, and it is what allows $B(\boldsymbol{k})$ to contribute more than a rescaling of $A(\boldsymbol{k})$. The four fields $(u_0^{\mathcal{I}}, v_0^{\mathcal{I}}, u_0^{\mathcal{E}}, v_0^{\mathcal{E}})$ are concatenated along the channel axis and transformed by a single batched rfft2, and the two outputs are returned by a single irfft2, which halves the number of cuFFT launches per block.

With $\lambda^m = \mathbf{0}$, equation (10) returns its input exactly, and the filter and coupling start close to the identity ($\Phi \approx 1$, $a = b \approx 1$, $\kappa \approx 0$, with perturbations of standard deviation 0.02). The residual feedback of equation (1) still changes the downstream encoder input at step 0. Because cuFFT has no half-precision path for the non-power-of-two sizes typical of driving feature maps (e.g. 50×87), the spectral part of the block runs in fp32 with autocast disabled, while the surrounding convolutions stay in fp16.

3.7 Decoder and Objective

The fused maps $\{F_i\}_{i=1}^4$ of equation (1) are decoded by the all-MLP head of SegFormer (Xie et al., 2021): per-level linear projection to a shared width, bilinear upsampling to $H/4 \times W/4$, concatenation, fusion by a 1×1 convolution and a linear classifier, followed by bilinear upsampling to the input resolution. Training minimises the pixel-wise cross-entropy over the set Ω of labelled pixels,

$$\mathcal{L} = -\frac{1}{|\Omega|}\sum_{p\in\Omega} \log \hat{y}_{p,c_p}, \tag{11}$$

where c_p is the ground-truth class of pixel p and $\hat{y}_{p,c_p}$ its predicted probability. We add no auxiliary loss, edge supervision or distillation term, so the reported gains are attributable to the fusion operator rather than to extra supervision.

4 Experiments

4.1 Datasets

DDD17 (Binas et al., 2017) is an end-to-end DAVIS driving dataset with paired greyscale APS frames and events at 346×260; we use the segmentation split with 6 classes introduced by Ev-

Table 1: Comparison on DERS-XS, DSEC-Semantic and DSEC-Xtrm. All competing numbers are taken from Bao et al. (2025), where every method is retrained under an identical data pipeline.

Method	Modality	DERS-XS		DSEC-Semantic		DSEC-Xtrm	
		mACC	mIoU	mACC	mIoU	mACC	mIoU
SegFormer (Xie et al., 2021)	RGB	62.45	53.21	72.91	65.03	41.74	33.88
SegFormer (E) (Xie et al., 2021)	Event	47.85	37.32	47.38	38.59	48.52	37.72
EvSegFormer (Jia et al., 2023)	Event	41.48	31.85	44.72	37.13	42.33	34.68
TokenFusion (Wang et al., 2022)	E-RGB	64.88	56.22	74.60	67.39	53.04	45.41
CMX (Zhang et al., 2023a)	E-RGB	71.86	63.12	76.18	68.10	51.29	43.95
CMNeXt (Zhang et al., 2023b)	E-RGB	73.30	64.55	77.50	69.03	52.12	45.16
EISNet[†] (Xie et al., 2024a)	E-RGB	69.10	60.68	71.60	64.67	56.77	48.76
ESC (Bao et al., 2025)	E-RGB	75.26	67.10	78.61	71.04	59.45	50.87
MambaFourier (ours)	E-RGB	**77.23**	**70.40**	**82.38**	**75.65**	**63.68**	**56.51**

[a] Pixel accuracy on DSEC-Semantic: 95.68%.

SegNet (Alonso & Murillo, 2019), with $15{,}950$ training and $3{,}890$ test pairs. **DSEC-Semantic** (Gehrig et al., 2021; Sun et al., 2022) provides 640×440 event–RGB pairs from a stereo event rig with 11-class labels ($8{,}082$ training / $2{,}809$ test frames). Its labels are pseudo-labels derived from RGB, which favours RGB-dominant models; we report it for comparability and interpret it accordingly. **DSEC-Xtrm** (Bao et al., 2025) degrades DSEC-Semantic to extreme conditions by unprocessing frames to the RAW domain, attenuating the optical signal and adding shot noise, and by injecting pure shot-noise events generated with v2e (Hu et al., 2021). Because the labels are inherited unchanged while both inputs are degraded, the gap between DSEC-Semantic and DSEC-Xtrm measures resilience directly. **DERS-XS** (Bao et al., 2025) is a true-labelled synthetic benchmark built with CARLA (Dosovitskiy et al., 2017) and v2e (Hu et al., 2021): 270 sequences, $2{,}016$ training / $1{,}080$ test pairs at 640×360, with low-light frames and noisy events and, crucially, ground-truth rather than RGB-derived labels. We report mean intersection-over-union (mIoU) and mean class accuracy (mACC).

4.2 Implementation and training details

Architecture. Both branches use GroupMamba-Small (Shaker et al., 2025) (dims $[64, 128, 348, 512]$, depths $[3, 4, 16, 3]$) initialised from ImageNet-1K weights; the event stem is adapted to B input channels by channel replication with $3/B$ energy rescaling. Stochastic depth is 0.2. WFTF uses $M = 16$ retained modes per axis at all four stages; at stage 4 the slices are clipped to the available frequencies. Coupling is enabled, out-of-band coefficients pass through, dynamic time and unitary normalisation are off by default.

Event representation. Voxel grids (Zhu et al., 2018) with $B = 5$ bins for DDD17/DERS-XS ($50\,\mathrm{ms}$ window) and $B = 15$ bins over $50{,}000$ events per sample for DSEC-Semantic/DSEC-Xtrm. AdamW (Loshchilov & Hutter, 2019) with initial learning rate 6×10^{-5} and weight decay 0.01, a polynomial decay schedule with power 0.9 stepped per optimiser update, and cross-entropy loss with the dataset's ignore index. We train for 100 epochs on DDD17 and DERS-XS and 60 epochs on DSEC-Semantic and DSEC-Xtrm. **Hardware and memory.** All models are trained on a single NVIDIA GeForce RTX 3090 (24 GB). **Augmentation.** Random horizontal flip, random resize with scale in $[0.5, 2.0]$ applied jointly to frames and voxel grids, colour jitter on the frame only, and random crop to 200×346 (DDD17) or 440×640 (DSEC). No test-time augmentation, no multi-scale inference and no model ensembling are used anywhere in this paper.

4.3 Comparison with the state of the art

Table 1 reports results on the true-labelled synthetic benchmark and on the two real-world DSEC variants; Table 2 reports DDD17.

DDD17. MambaFourier reaches 79.39% mIoU at 96.50% pixel accuracy, improving on MambaSeg (Gu et al., 2026) by 1.83% mIoU and on EISNet (Xie et al., 2024a) by 4.36%. The compar-

Table 2: Comparison with RGB- and event-based semantic segmentation methods on DDD17.

Method	Modality	Backbone	Event rep.	mIoU (%)	Acc. (%)
SegFormer (Xie et al., 2021)	RGB	MiT-B2	–	71.05	95.73
CMESS (Xie et al., 2024b)	RGB-Event	E2VID	Voxel grid	64.30	92.07
OpenESS (Kong et al., 2024)	RGB-Event	E2VID	Voxel grid	63.00	91.05
SpikingEDN (Zhang et al., 2025)	RGB-Event	FCN	Voxel grid	72.57	–
SAM-Event-Adapter (Yao et al., 2024)	RGB-Event	SAM	MSP	69.06	95.32
EISNet (Xie et al., 2024a)	RGB-Event	MiT-B2	AET	75.03	96.04
Spike-BRGNet (Long et al., 2025)	RGB-Event	MiT-B2	Voxel grid	54.72	–
Hybrid-Seg (Li et al., 2025)	RGB-Event	FCN	Voxel grid	67.31	95.07
MambaSeg (Gu et al., 2026)	RGB-Event	Mamba	Voxel grid	77.56	96.31
MambaFourier (ours)	RGB-Event	Mamba	Voxel grid	**79.39**	**96.50**

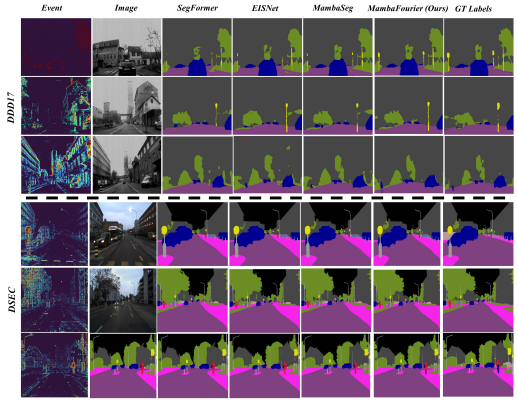


Figure 4: Qualitative comparison on DDD17 (top) and DSEC-Semantic (bottom).

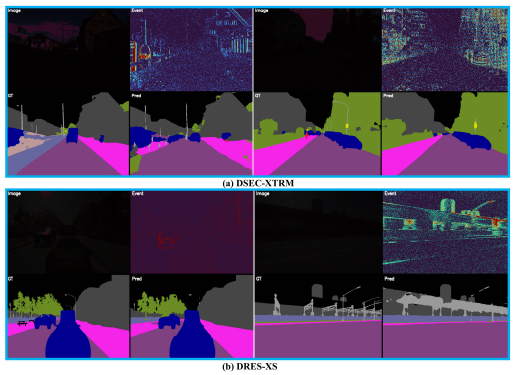


Figure 5: Qualitative comparison under extreme conditions on DSEC-Xtrm (top) and DERS-XS (bottom)..

ison with MambaSeg is the informative one: both use a Mamba backbone and the same voxel-grid representation, so the margin isolates the contribution of the fusion operator rather than of the encoder family.

DSEC-Semantic and DSEC-Xtrm. We obtain 75.65% mIoU, 4.61% above ESC (Bao et al., 2025) and 6.62% above CMNeXt (Zhang et al., 2023b). This benchmark's labels are RGB-derived pseudo-labels, so it structurally rewards RGB-dominant behaviour; it is therefore notable that a method designed around event-borne high-frequency evidence still leads by a clear margin, and it suggests that the propagator is contributing spatial regularisation to the RGB branch in addition to cross-modal transfer. In DSEC-Xtrm dataset the margin widens to 5.64% mIoU over ESC and 11.35% over CMNeXt. The relative degradation from DSEC-Semantic to DSEC-Xtrm is the cleanest resilience measure available, since only the inputs change: CMNeXt loses 23.87 mIoU points, ESC loses 20.17, and MambaFourier loses 19.14 while starting from a far higher baseline.

DERS-XS. On true-labelled synthetic data we obtain 70.40% mIoU, 3.30% above ESC and 5.85% above CMNeXt. Because these labels are generated by CARLA rather than inferred from RGB, this is the benchmark least biased toward either modality, and the gain there is the one we regard as most diagnostic.

Figures 4 and 5 compare predictions qualitatively. Under normal illumination (Figure 4) the difference is concentrated on thin and small-area classes, which is consistent with a fusion operator that adds selective high-frequency capacity.

Table 3: Ablation on DDD17. Acc. denotes pixel accuracy.

Variant	mIoU (%)	Acc. (%)
Fusion block; shared encoder and decoder		
EISNet fusion	75.01	95.96
MRFM	76.58	95.97
CSF	76.65	96.22
FFM	76.77	96.16
DDIM	77.56	96.23
SWP (DCT wave gating)	78.75	–
WFTF component; one change at a time		
No coupling	77.78	95.72
No propagator	77.93	95.75
No free-form filter	77.68	95.71
Hard band limit	78.35	95.86
$M = 8$ retained modes	78.82	96.30
Full WFTF	**79.39**	**96.50**

Table 4: Model complexity evaluated on DSEC and DRES-XS Datasets.

Method	Backbone	#Params (M)	FLOPs (G)
SegFormer	MiT-B2	24.725	25.279
SegFormer (E)	MiT-B2	24.734	25.433
EvSegFormer	MiT-B2	24.740	25.422
SwiftFormer	Transformer	64.61	23.88
TokenFusion	MiT-B2	26.011	54.845
CMX	2×MiT-B2	66.566	65.551
CMNeXt	2×MiT-B2	58.687	62.805
EISNet	MiT-B0 + B2	34.367	67.304
ESCN	MiT-B1 + B2	56.875	95.086
MambaSeg	VMamba-T	97.55	71.01
MambaFourier (ours)	GroupMamba-S	**74.13**	**65.04**

4.4 Model complexity

Table 4 compares parameter counts and computational cost. MambaFourier uses 74.13M parameters and 65.04G FLOPs, compared with 97.55M parameters and 71.01G FLOPs for MambaSeg. This corresponds to 24.0% fewer parameters and 8.4% fewer FLOPs, alongside the 1.83-point improvement in DDD17 mIoU reported in Table 2. The methods use different backbones and input configurations, so their costs should be interpreted alongside segmentation accuracy rather than as an architecture-independent ranking.

4.5 Ablation study

The upper part of Table 3 isolates the fusion block: each variant uses the same dual-GroupMamba encoder, all-MLP decoder and training recipe. Full WFTF achieves 79.39% mIoU, exceeding DDIM, the strongest non-spectral learned fusion variant, by 1.83 points. DCT-based wave gating (SWP) reaches 78.75% mIoU, 0.64 points below WFTF. The bottom block changes one WFTF component at a time. Removing the cross-modal coupling ($\kappa \equiv 0$) is the most damaging change. It keeps per-modality propagation and filtering but prevents any spectral exchange, and mIoU falls from 79.39 to 77.78, only 0.22 above DDIM. Almost all of the gain over the strongest learned module therefore depends on mode-wise exchange rather than on added capacity. Removing either spectral operator costs more than 1.4 mIoU. The free-form filter alone (77.93, a GFNet-style filter (Rao et al., 2021)) is slightly stronger than the three-parameter propagator alone (77.68), but together they reach 79.39. The physically parameterised response and the unconstrained spectral weights thus capture complementary structure. Zeroing out-of-band coefficients turns WFTF into a hard low-pass operator and costs 1.04 mIoU, consistent with the loss of thin structures discussed in Section 3.4. Halving the retained band to $M = 8$ costs only 0.57, so the block is not sensitive to the exact width of the band.

5 Conclusion

We introduced **MambaFourier**, a dual-branch state-space segmentation network whose cross-modal fusion is physics-derived. The Wave–Fourier Fusion block treats paired RGB and event features as coupled damped wave fields and evaluates their interaction with the exact closed-form solution of the damped wave equation in the Fourier basis, augmented by a band-limited free-form filter and a reciprocal per-frequency coupling matrix. MambaFourier targets a critical low-light segmentation failure: insufficient illumination destroys RGB texture, making appearance-based features unreliable, while events preserve high-frequency structure but are sparse and intensity-independent. MambaFourier achieves state-of-the-art results: 79.39% mIoU on DDD17 ($+1.83\%$ over MambaSeg), 75.65% on DSEC-Semantic ($+4.61\%$ over ESC), 70.40% on DERS-XS ($+3.30\%$ over ESC), and 56.51% on DSEC-Xtrm ($+5.64\%$ over the previous best). Under synthetic rain and fog on DSEC, it reaches 73.7% mIoU versus 72.5% for MambaSeg, while using fewer parameters and FLOPs.

AI USE STATEMENT

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

In this work, we used generative AI tools for language polishing of the manuscript and for assistance with boilerplate code and LaTeX formatting. We have not used generative AI tools for research ideation, experimental design, derivation of the analytical results, or generation of experimental data or figures. Retrieval of related work was performed by the authors; AI tools were not used as the sole source for any citation, and every reference in this paper was verified against its original publication venue by the authors. All AI-assisted text and code were reviewed, edited and tested by the authors. We take responsibility for the final content of this work, including all text, claims and artifacts.

ETHICS STATEMENT

This work uses only publicly available driving datasets (DDD17, DSEC-Semantic) and simulated data (DERS-XS, DSEC-Xtrm) that contain no personally identifying annotations; no human subjects were involved. We note a dual-use consideration common to perception research for autonomous driving: improved low-light segmentation can support safety systems, but deployment claims should not be inferred from benchmark performance, and the failure case identified in Section 4.4 (no motion and no illumination) is a real operational limit rather than a residual error.

REPRODUCIBILITY STATEMENT

The architecture is specified in Section 3, with the propagator in equation (6) and the coupling in equation (8). All optimiser settings, schedules, event representations, augmentation policy, batch sizes, accumulation steps and hardware are given in Section 4.2; every result in this paper was produced on a single NVIDIA RTX 3090 without test-time augmentation or ensembling. Dataset construction for DERS-XS and DSEC-Xtrm follows Bao et al. (2025), whose released pipelines we use unmodified. Anonymised source code, including the self-test suite that verifies identity-at-initialisation, resolution independence and gradient flow through every spectral parameter, is provided as supplementary material.

ACKNOWLEDGMENTS

To be added for the camera-ready version.

REFERENCES

Iñigo Alonso and Ana C. Murillo. EV-SegNet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

Nan Bao, Yifan Zhao, Lin Zhu, and Jia Li. Re-coding for uncertainties: Edge-awareness semantic concordance for resilient event-RGB segmentation. In *Advances in Neural Information Processing Systems*, 2025.

Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. DDD17: End-to-end DAVIS driving dataset. *arXiv preprint arXiv:1711.01458*, 2017.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16, 2017.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022.

Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Fuqiang Gu, Yuanke Li, Xianlei Long, Kangping Ji, Chao Chen, Qingyi Gu, and Zhenliang Ni. MambaSeg: Harnessing Mamba for accurate and efficient image-event semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4302–4310, 2026.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Efficient token mixing for transformers via adaptive Fourier neural operators. In *International Conference on Learning Representations*, 2022.

Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. SegNeXt: Rethinking convolutional attention design for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1140–1156, 2022.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic DVS events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1312–1321, 2021.

Zexi Jia, Kaichao You, Weihua He, Yang Tian, Yongxiang Feng, Yaoyuan Wang, Xu Jia, Yihang Lou, Jingyi Zhang, Guoqi Li, and Ziyang Zhang. Event-based semantic segmentation with posterior attention. *IEEE Transactions on Image Processing*, 32:1829–1842, 2023.

Lingdong Kong, Youquan Liu, Lai Xing Ng, Benoit R. Cottereau, and Wei Tsang Ooi. OpenESS: Event-based semantic scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15686–15698, 2024.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. FNet: Mixing tokens with Fourier transforms. In *Proceedings of NAACL-HLT*, pp. 4296–4313, 2022.

Hebei Li, Yansong Peng, Jiahui Yuan, Peixi Wu, Jin Wang, Yueyi Zhang, and Xiaoyan Sun. Efficient event-based semantic segmentation via exploiting frame-event fusion: A hybrid neural network approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18296–18304, 2025.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

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

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103031–103063, 2024.

Xianlei Long, Xiaxin Zhu, Fangming Guo, Chao Chen, Xiangwei Zhu, Fuqiang Gu, Songyu Yuan, and Chunlong Zhang. Spike-BRGNet: Efficient and accurate event-based semantic segmentation with boundary region-guided spiking neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3):2712–2724, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 980–993, 2021.

Abdelrahman Shaker, Syed Talal Wasim, Salman Khan, Juergen Gall, and Fahad Shahbaz Khan. GroupMamba: Efficient group-based visual state space model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. ESS: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pp. 341–357, 2022.

Hans Thisanke, Chamli Deshan, Kavindu Chamith, Sachith Seneviratne, Rajith Vidanaarachchi, and Damith Herath. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126:106669, 2023.

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12186–12195, 2022.

Bochen Xie, Yongjian Deng, Zhanpeng Shao, and Youfu Li. EISNet: A multi-modal fusion network for semantic segmentation with events and images. *IEEE Transactions on Multimedia*, 26:8639–8650, 2024a.

Chuyun Xie, Wei Gao, and Ren Guo. Cross-modal learning for event-based semantic segmentation via attention soft alignment. *IEEE Robotics and Automation Letters*, 9(3):2359–2366, 2024b.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

Bowen Yao, Yongjian Deng, Yuhan Liu, Hao Chen, Youfu Li, and Zhen Yang. SAM-Event-Adapter: Adapting segment anything model for event-RGB semantic segmentation. In *IEEE International Conference on Robotics and Automation*, pp. 9093–9100, 2024.

Zhen Yao, Xiaowen Ying, Zhiyu Zhu, and Mooi Choo Chuah. Learning flow-guided registration for RGB–event semantic segmentation. *arXiv preprint arXiv:2505.01548*, 2025.

Bo-Wen Yin, Jiao-Long Cao, Xuying Zhang, Yuming Chen, Ming-Ming Cheng, and Qibin Hou. OmniSegmentor: A flexible multi-modal learning framework for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24(12):14679–14694, 2023a.

Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1136–1147, 2023b.

Rui Zhang, Luziwei Leng, Kaiwei Che, Hu Zhang, Jie Cheng, Qinghai Guo, Jianxing Liao, and Ran Cheng. Accurate and efficient event-based semantic segmentation using adaptive spiking encoder–decoder network. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5): 9326–9340, 2025.

Yucheng Zhao, Gengyu Lyu, Ke Li, Zihao Wang, Hao Chen, Zhen Yang, and Yongjian Deng. ESEG: Event-based segmentation boosted by explicit edge-semantic guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10510–10518, 2025.

Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *Proceedings of the European Conference on Computer Vision Workshops*, pp. 711–714, 2018.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 62429–62442, 2024.

A Appendix

B GroupMamba backbone

We use independent GroupMamba-Small encoders for the image and event streams, with stage depths $[3, 4, 16, 3]$ and channel widths $[64, 128, 348, 512]$. In each GroupMamba layer, channels are divided into four groups processed by selective scans in complementary spatial directions; channel modulation combines information across the groups (Shaker et al., 2025). Both encoders are initialised from ImageNet-1K pretrained weights. For the event encoder, we adapt the three-channel stem weights to T voxel-grid channels by replication and $3/T$ scaling. We extract a feature map after each stage and apply WFTF to the paired image and event features before passing them to the decoder and subsequent stages.

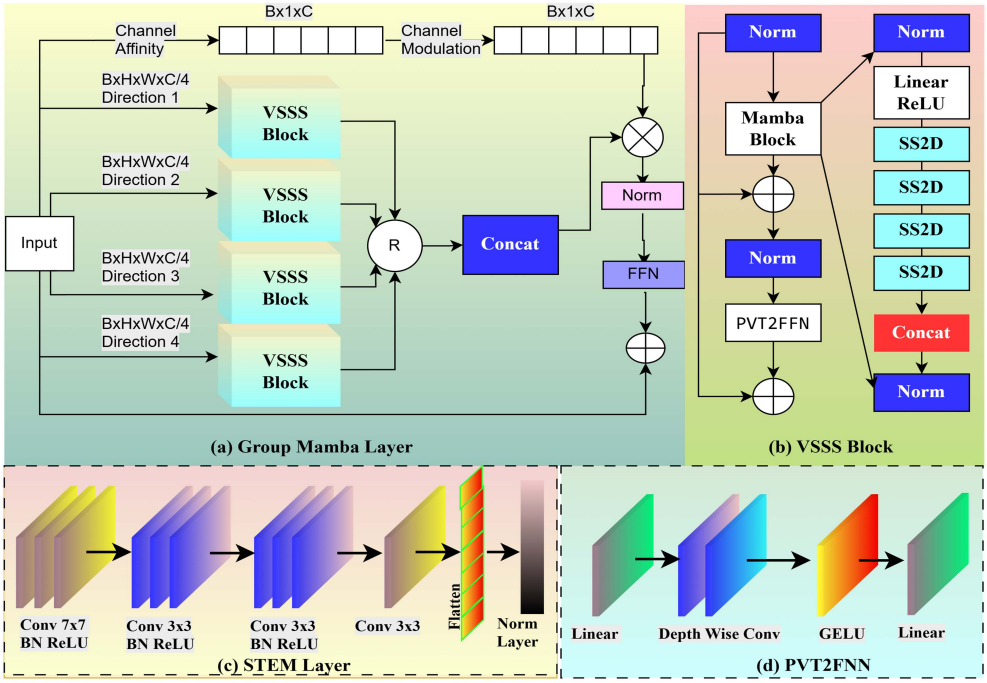


Figure 6: GroupMamba components used in each modality encoder (Shaker et al., 2025). The schematic shows the grouped directional scans, VSSS block, stem, and feed-forward module. Our WFTF blocks operate between encoder stages and are not shown.

C Technical details of the WFTF block

C.1 Exactness of the propagator

For each wavenumber, equation (3) has characteristic polynomial $s^2 + 2\alpha_{\boldsymbol{k}} s + \omega_{\boldsymbol{k}}^2 + \alpha_{\boldsymbol{k}}^2$ with roots $s = -\alpha_{\boldsymbol{k}} \pm i\,\omega_{\boldsymbol{k}}$. The general solution is therefore $\hat{u}(\tau) = e^{-\alpha_{\boldsymbol{k}}\tau}\big(C_1 \cos \omega_{\boldsymbol{k}}\tau + C_2 \sin \omega_{\boldsymbol{k}}\tau\big)$. The initial conditions give $C_1 = \hat{u}_0$ and $-\alpha_{\boldsymbol{k}}\hat{u}_0 + \omega_{\boldsymbol{k}} C_2 = \hat{v}_0$, which yields equation (4). The biharmonic term in equation (2) is what makes the discriminant negative for every $\boldsymbol{k}$, c and γ. Without it, modes with $\gamma\|\boldsymbol{k}\| > c$ become overdamped and equation (4) no longer applies.

The implemented update, equation (6), uses the offset $\gamma\hat{U}$ in place of $\alpha_{\boldsymbol{k}}\hat{U}$. Substituting $\hat{v}_0 = \hat{V} + (\gamma - \alpha_{\boldsymbol{k}})\hat{U}$ into equation (4) recovers equation (6) exactly. The update is thus the exact solution for this effective initial velocity. Both multipliers are bounded, $|A(\boldsymbol{k})| \le 1$ and $|B(\boldsymbol{k})| \le \tau$. Every propagated mode therefore satisfies $|\hat{P}^m(\boldsymbol{k})| \le (1 + \gamma\tau)\,|\hat{U}^m(\boldsymbol{k})| + \tau\,|\hat{V}^m(\boldsymbol{k})|$, so the propagator

cannot amplify any mode without bound. We compute B as $\tau\,\mathrm{sinc}(c\|\boldsymbol{k}\|\tau/\pi)$, so no division occurs at $\boldsymbol{k} = \mathbf{0}$, where $B = \tau$ exactly.

C.2 Initialisation and hyperparameters

The propagator's raw parameters are initialised to $c_{\mathrm{raw}} = \tau_{\mathrm{raw}} = 0.5413$ and $\gamma_{\mathrm{raw}} = -2$, giving $c = \tau = 1.00$ and $\gamma = 0.127$ after the softplus. The free-form filter and the coupling diagonal are initialised to $1 + 0j$, and the coupling off-diagonal to 0. All of these receive Gaussian perturbations with standard deviation 0.02 on both real and imaginary parts. LayerScale λ^m is initialised to exactly zero, which makes the block an exact identity at step 0 despite the perturbations. We use $M = 16$ at all four stages, pass-through outside the band, and disable both dynamic time and the coupling normalisation. The spectral section runs in fp32 inside `autocast(enabled=False)`. All transforms use orthonormal scaling (`norm="ortho"`).

C.3 Retained-mode coverage

Because parameters are indexed by mode number, the fraction of the spectrum they control depends on the feature-map size. Table 5 lists the retained band (m_h, m_w) and the share of rfft2 coefficients that carry learnable filter and coupling weights ($2m_h m_w$ of $H_i(\lfloor W_i/2 \rfloor + 1)$). At stage 1 on DSEC, only 5.7% of the coefficients are learnable, while the propagator still acts on all of them. By stage 4 the band covers the whole spectrum. The same 16×16 weight blocks serve both datasets; only the slice that is read changes.

Table 5: Retained modes per stage with $M = 16$ for the training crops of DDD17 (200×346) and DSEC (440×640).

	DDD17				**DSEC**			
Stage	$H_i \times W_i$	m_h	m_w	learnable	$H_i \times W_i$	m_h	m_w	learnable
1	50×87	16	16	23.3%	110×160	16	16	5.7%
2	25×44	12	16	66.8%	55×80	16	16	22.7%
3	13×22	6	12	92.3%	28×40	14	16	76.2%
4	7×11	3	6	85.7%	14×20	7	11	100.0%

C.4 Parameter budget

Table 8 counts the parameters of each WFTF block and of the 1×1 reduction convolution that forms F_i. The counts are exact for the released implementation with $M = 16$. The spectral parameters are $20CM^2 + 6C$ per block ($8CM^2$ for the two filters, $12CM^2$ for the coupling and $3C$ per propagator); the rest are the normalisation, depthwise, projection and output convolutions. The four stages together add 8.66M parameters to the two encoders.

Table 6: Parameters of the fusion stage (in millions), GroupMamba-Small, $M = 16$.

Stage	C_i	Spectral	Convolutions	Reduction 1×1	Total
1	64	0.328	0.027	0.008	0.363
2	128	0.656	0.102	0.033	0.791
3	348	1.784	0.738	0.243	2.764
4	512	2.625	1.589	0.525	4.739
All	–	5.393	2.456	0.808	8.657

C.5 Runtime, throughput and memory

We measure training cost on a single NVIDIA RTX 3090 (24 GB) on DDD17 with batch size 16 and 997 iterations per epoch, i.e. $15{,}952$ event–image pairs per epoch. Peak allocated memory is

reported in fp32 at 200×346 with batch size 1. Table 7 summarises the results. MambaFourier reduces the epoch time from 9 min 50 s to 8 min, i.e. by 18.6%, and increases training throughput from 27.0 to 33.2 img/s, i.e. by 22.9%. It also uses 24.0% fewer parameters (74.14M vs. 97.55M), and its checkpoint is 23.6% smaller (298 MB vs. 390 MB). At inference, peak allocated memory drops by 19.3% (0.421 GB vs. 0.522 GB).

Table 7: Training cost, parameter count, peak allocated memory, and checkpoint size on an RTX 3090, DDD17.

Method	**Params** (M)	**Fusion** (M)	**Fusion** share (%)	**Epoch** (min)	**Train** (img/s)	**Peak alloc.** (GB)	**Ckpt** (MB)
MambaSeg	97.55	33.54	34.4	9.83	27.0	0.522	390
MambaFourier	**74.14**	**7.85**	**10.6**	**8.00**	**33.2**	**0.421**	**298**

The parameter breakdown in Table 8 shows that the fusion module accounts for only 7.85M parameters in MambaFourier, i.e. 10.6% of the total, compared with 33.54M (34.4%) in MambaSeg (Gu et al., 2026). This is far below the $\sim$ 28M fusion parameter scale sometimes attributed to the fusion block.

Table 8: Parameter breakdown (M) on the RTX 3090.

Component	**MambaSeg**	**MambaFourier**
Image encoder	29.48	31.14
Event encoder	29.48	31.16
Fusion (CCBS, MWCMF $\times 4$)	33.54	7.85
Output convolutions	1.57	0.81
Decoder	3.48	3.18
Total	**97.55**	**74.14**

D Robustness and additional qualitative results

D.1 Rain and fog

We corrupt the RGB stream of DSEC with synthetic rain and fog and leave the events untouched (Figure 7). All models are trained with the same corruption augmentation; both baselines are retrained by us (†). MambaFourier attains $\mathbf{73.7}$ mIoU, compared with $72.5^{\dagger}$ for MambaSeg and $71.5^{\dagger}$ for EISNet. The two corruptions damage different parts of the image spectrum. Rain streaks add high-wavenumber structure that is not tied to scene geometry, and the Kelvin–Voigt envelope $e^{-\gamma\|\boldsymbol{k}\|^2\tau}$ attenuates exactly this band. Fog lowers contrast and flattens the low-frequency layout that the frame normally contributes. Because the coupling $\kappa(\boldsymbol{k})$ is indexed by wavenumber, the event stream, which neither corruption affects, can supply structure mode by mode rather than through a single spatial gate. We attribute the margin over attention- and gating-based fusion to this band-selective exchange. We have not isolated it experimentally under corruption.

D.2 Simulated dark night

To probe low illumination beyond DSEC-Xtrm, we use the ESC protocol to darken both the training and the test sequences of DSEC to simulate night-time capture. Events are left unchanged, since an event camera responds to relative rather than absolute brightness change. Figure 8 shows that MambaFourier keeps the road surface coherent and still separates vehicles, poles and traffic signs where the darkened frame shows almost no texture. In these regions the prediction is supported mainly by the event stream, which reaches the image branch through the coupling $\kappa(\boldsymbol{k})$ and the residual fusion feedback of equation (1).

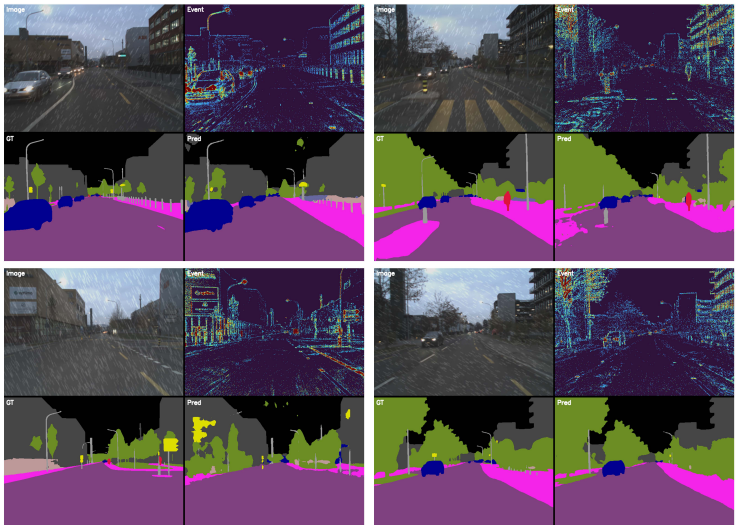


Figure 7: Qualitative results on DSEC under synthetic rain and fog applied to the RGB stream only. Events are unaltered. MambaFourier preserves road layout and object boundaries where the corrupted frame alone is unreliable.

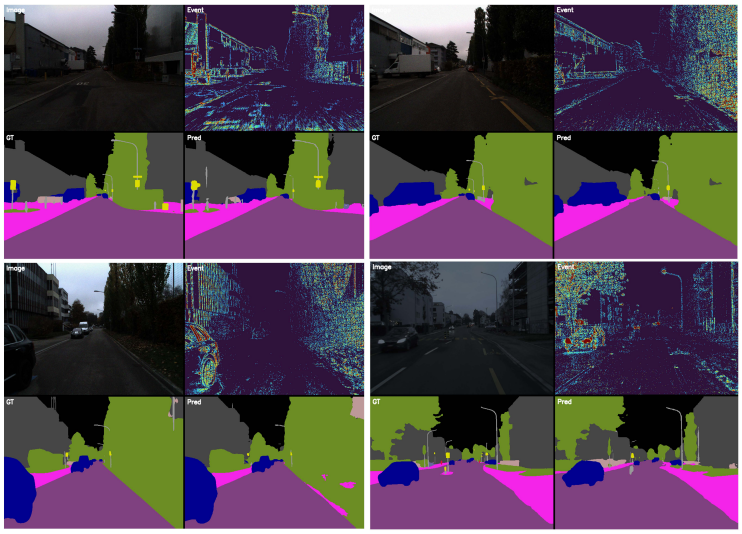


Figure 8: Qualitative results on DSEC with simulated dark-night conditions, applied to both the training and the test sequences. The RGB frame retains almost no texture, while the prediction preserves the road layout and thin structures.

D.3 Detailed results on DSEC-Xtrm

Figure 9 gives a detailed view of DSEC-Xtrm (Bao et al., 2025). Each composite shows the degraded frame, the event representation, the ground truth and our prediction. Because the degraded frames are close to black, we raise the brightness of the image panels *for display only*. The network receives the original degraded frames, and no enhancement is applied at training or test time. The brightened panels make visible how little usable structure the frame retains and how much of the predicted layout must come from the events.

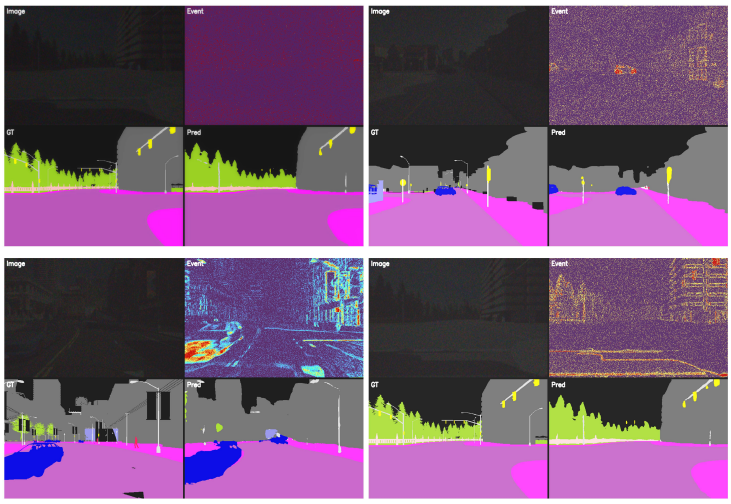


Figure 9: Detailed qualitative results on DSEC-Xtrm. Each composite shows, from left to right, the degraded RGB frame, the events, the ground truth and the MambaFourier prediction. Image panels are brightened for visualisation only; the model sees the unmodified degraded input.

D.4 Self-collected campus driving dataset

To test the method outside public benchmarks, we recorded a campus driving dataset following the DSEC acquisition and annotation protocol. It contains 6,856 training and 1,083 test samples, each consisting of an RGB frame, the synchronised events and a dense semantic label map. Trained with the recipe of Section 4.2, MambaFourier reaches 68.1% mIoU and 95.95% pixel accuracy on the test split. Figure 10 shows examples. The scenes differ from DSEC in layout and object statistics, including narrow roads, dense vegetation, pedestrians and cyclists, and the predictions keep coherent road and structure boundaries under these conditions.

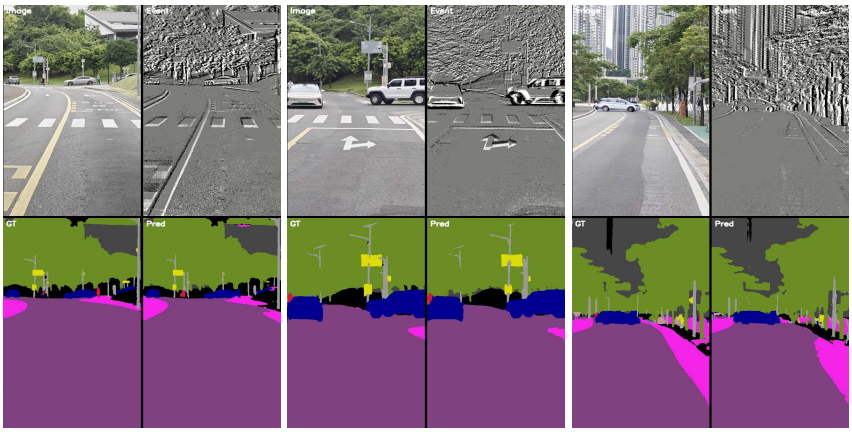


Figure 10: Qualitative results on our self-collected campus driving dataset (6,856 training / 1,083 test samples, recorded following the DSEC protocol). Each composite shows the RGB frame, the events, the ground truth and the MambaFourier prediction. MambaFourier attains 68.1% mIoU and 95.95% pixel accuracy on the test split.