WaveMamba: Wave-Inspired Cross-Modal Fusion for Robust event-image Semantic Segmentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Event-image semantic segmentation is promising for autonomous driving and robotics because event cameras complement frame images under rapid motion, low illumination, and high dynamic range. However, effective multimodal segmentation remains difficult due to asynchronous sensing, sparse-dense modality mismatch, and the cost of heavy cross-modal interaction modules. We propose **WaveMamba**, a dual-branch event-image semantic segmentation framework that combines hierarchical **GroupMamba** encoders with a **Mamba-Wave Cross-Modal Fusion (MWCMF)** module. Instead of relying on explicit registration or attention-heavy fusion, MWCMF performs implicit cross-modal propagation through channel-wise and spatial-wise wave-guided gating in the spectral domain, followed by Mamba-based refinement. Experiments on DDD17 and DSEC show that WaveMamba achieves state-of-the-art **78.75** and **76.00** mIoU, respectively. Under rain- and fog-corrupted evaluation on DSEC, WaveMamba further attains **74.0** mIoU, indicating improved robustness under degraded visual conditions.

1 Introduction

Semantic segmentation is a core vision task in autonomous driving and robotics, where a model assigns a semantic label to each pixel [1, 2]. Recent RGB-based models have achieved strong performance with convolutional and transformer architectures [3, 4], but frame cameras remain vulnerable to rapid motion, motion blur, low illumination, and large dynamic range variation [5, 6]. These limitations are especially problematic in safety-critical settings.

Event cameras provide a complementary sensing modality by recording asynchronous brightness changes with microsecond-level temporal resolution, low latency, and wide dynamic range [7, 8]. However, event streams are sparse, edge-dominant, and lack dense semantic appearance cues [9, 10]. This has motivated growing interest in *event-image semantic segmentation*, which aims at integrating intensive image semantics with event-based motion data.

A key challenge in event-image segmentation is *cross-modal inconsistency*. Images provide dense appearance and texture, whereas events emphasize motion boundaries and temporal change. The two modalities also differ in sensing format and temporal granularity, creating both representation mismatch and spatiotemporal inconsistency. Existing methods mainly address this challenge through direct fusion, such as cross-attention, gated interaction, or feature rectification [11, 12, 13, 14]. Although effective, these designs can introduce substantial interaction cost and may still struggle when the modalities are weakly aligned. Other approaches use explicit correspondence modeling or alignment heuristics, which can improve consistency but add complexity and dependence on external motion cues.

State-space models offer an efficient alternative to attention for dense multimodal prediction. Recent advances such as S4, Mamba, and Vision Mamba show that selective state-space modeling can

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

capture long-range dependencies efficiently [15, 16, 17]. In event-image segmentation, MambaSeg demonstrates that dual-branch Mamba encoders with explicit channel-spatial interaction can achieve strong performance [18]. However, existing Mamba-based fusion designs still rely largely on direct interaction modules and do not explicitly view cross-modal exchange as a propagation process.

In this paper, we propose **WaveMamba**, a dual-branch event-image semantic segmentation framework that combines parallel **GroupMamba** encoders with a new **Mamba-Wave Cross-Modal Fusion (MWCMF)** module. The two GroupMamba branches independently encode image and event inputs into hierarchical multiscale representations, providing efficient long-range dependency modeling and strong contextual aggregation. To improve cross-modal interaction, MWCMF introduces two complementary pathways: **Cross-Modal Channel-Wave Gating (CCWG)** and **Cross-Modal Spatial-Wave Gating (CSWG)**. Rather than relying on heavy cross-attention or explicit pixel-wise registration, MWCMF performs implicit cross-modal propagation in the spectral domain, followed by Mamba-based refinement along channel-wise and spatial-wise paths. This design encourages complementary information exchange while limiting noisy interference between dense image semantics and sparse event dynamics.

Our contributions are summarized as follows:

- We propose **WaveMamba**, a dual-branch event-image semantic segmentation framework that combines efficient **GroupMamba** encoders with multiscale cross-modal fusion.
- We introduce **MWCMF**, a fusion module that performs implicit cross-modal propagation through **channel-wave** and **spatial-wave** gating, followed by Mamba-based refinement.
- We show through controlled ablations that both the **GroupMamba** backbone and the proposed **MWCMF** contribute to the final gains over strong Mamba-based baselines.
- We conduct extensive experiments on DDD17 and DSEC, where WaveMamba achieves **78.75 mIoU** and **76.0 mIoU**, respectively. We further investigate robustness under rain and fog corruptions and obtain **74.0 mIoU** on the augmented DSEC validation setting, substantially outperforming a reproduced MambaSeg baseline.

2 Related Work

2.1 Event-image semantic segmentation

Semantic segmentation with event cameras has progressed from event-only recognition to multimodal event-image perception. Early methods showed that event data can support scene understanding, but the lack of dense appearance cues limited performance on complex urban scenes [9, 19]. Later methods therefore fused event streams with frame images to exploit both dense semantics and motion cues.

Most prior work follows a *fusion-centric* paradigm. CNN- and transformer-based methods introduce cross-modal rectification, gated interaction, progressive recalibration, and hybrid ANN-SNN designs to combine image and event representations [11, 12, 13, 20, 21, 22, 23]. These methods have improved performance on DDD17 and DSEC, but many rely on relatively heavy interaction modules or heuristic alignment mechanisms. In contrast, our work explores a lighter implicit propagation mechanism in the spectral domain.

2.2 Event representation

A central question in event-based perception is how to represent asynchronous event streams for dense prediction. Common choices include image-like accumulations, polarity-separated maps, and voxel grids obtained by temporal binning [24, 25, 26, 27]. Among these, voxel grids are widely used because they provide a practical bridge between asynchronous events and dense neural architectures [10, 20]. Recent methods have also proposed refined encodings that emphasize scene activity or multiscale temporal structure [13, 22].

While these representations improve compatibility with standard backbones, they do not fully resolve the mismatch between sparse event responses and texture-rich image features. Our work therefore keeps a dense-compatible event representation and focuses on the interaction stage.

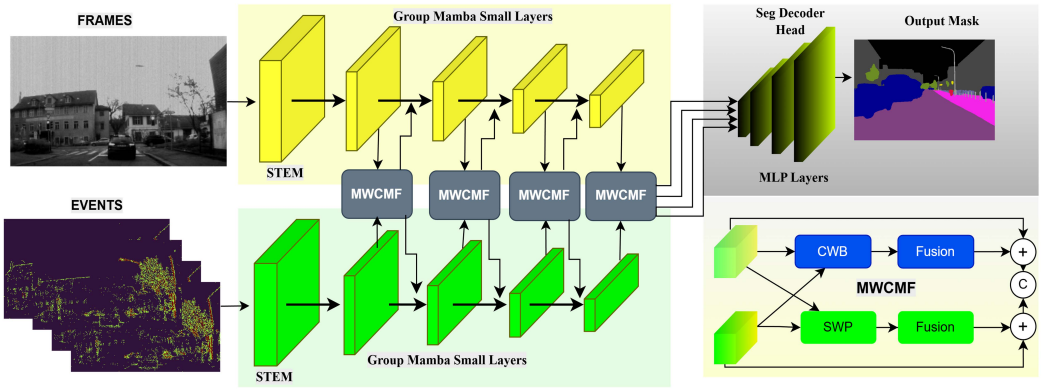

Figure 1: Overview of WaveMamba. Image and event inputs are encoded by parallel GroupMamba branches. At each scale, MWCMF performs bidirectional cross-modal interaction through channel-wave and spatial-wave pathways, and the resulting multiscale fused features are decoded by a lightweight MLP segmentation head.

2.3 State-space models and cross-modal alignment

State-space models have recently emerged as an efficient alternative to attention for long-range sequence modeling. S4 established the promise of structured state-space layers, and Mamba introduced selective state-space modeling with linear-time complexity and strong empirical performance [15, 16]. Vision Mamba and related variants have since been adapted to classification and dense prediction [17, 28, 29].

In multimodal event-image segmentation, MambaSeg shows that a dual-branch Mamba architecture can combine efficient representation learning with explicit channel-spatial interaction [18]. Our work shares the motivation of using efficient state-space modeling, but differs in both backbone and fusion design. WaveMamba adopts **GroupMamba** encoders for hierarchical multiscale representation learning and introduces **MWCMF**, which performs implicit cross-modal propagation through wave-guided channel-wise and spatial-wise gating. This differs from both heavier direct fusion and explicit registration-based alignment approaches.

3 Methodology

3.1 Overview

We propose **WaveMamba**, a dual-branch event-image semantic segmentation framework that combines parallel **GroupMamba** encoders with a **Mamba-Wave Cross-Modal Fusion (MWCMF)** module. Given an image frame and its synchronized event representation, WaveMamba first extracts hierarchical features from the two modalities independently and then performs cross-modal interaction at each scale through MWCMF. The resulting multiscale fused features are decoded by a lightweight SegFormer-style MLP head to predict the final segmentation map. An overview of the framework is shown in Fig. 1, while the designs of MWCMF and the GroupMamba blocks are illustrated in Fig. 2 and Fig. 3, respectively.

Let the image input be denoted by

$$I \in \mathbb{R}^{C_I \times H \times W}, \tag{1}$$

and let the event stream be converted into a voxel-grid representation

$$E \in \mathbb{R}^{C_E \times H \times W}, \tag{2}$$

where C_I and C_E denote the numbers of image and event channels. In our setting, the image branch receives grayscale or RGB frames, while the event branch uses temporally binned event voxels.

WaveMamba applies a modality-specific stem layer followed by four hierarchical GroupMamba stages. At the s-th stage, the image and event features are denoted by

$$F_I^{(s)} \in \mathbb{R}^{C_s \times H_s \times W_s}, \qquad F_E^{(s)} \in \mathbb{R}^{C_s \times H_s \times W_s}. \tag{3}$$

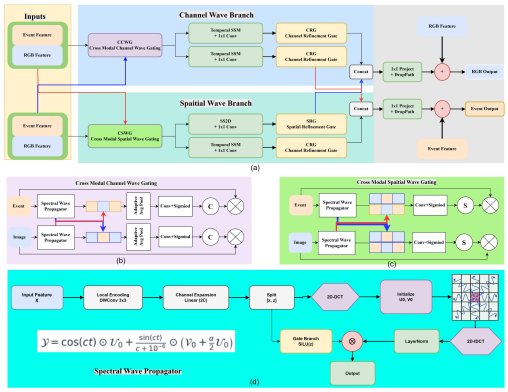

Figure 2: Details of the proposed MWCMF module. (a) Overall workflow of MWCMF (b) CCWG (c) CSWG (d) SWP

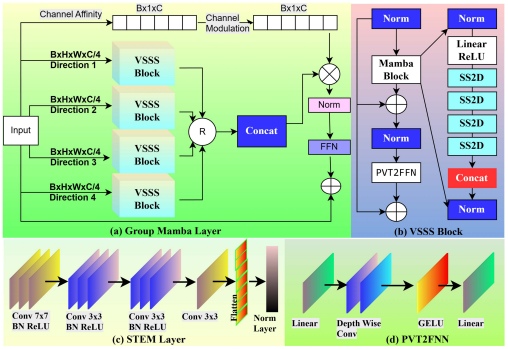

Figure 3: GroupMamba building blocks used in WaveMamba. (a) GroupMamba layer. (b) VSSS block. (c) Stem layer. (d) PVT2FFN.

These features are passed to MWCMF for bidirectional cross-modal refinement:

$$\tilde{F}_I^{(s)}, \tilde{F}_E^{(s)} = \mathrm{MWCMF}\left(F_I^{(s)}, F_E^{(s)}\right). \tag{4}$$

The refined outputs are propagated to the next encoder stage, and the fused multiscale representations are collected for decoding.

3.2 Input representation and hierarchical encoding

Event cameras asynchronously record brightness changes as a stream of events

$$e_k = (x_k, y_k, t_k, p_k), \tag{5}$$

where (x_k, y_k) denotes the pixel coordinate, t_k the timestamp, and $p_k \in \{-1, +1\}$ the event polarity. Following standard practice, we discretize the event stream into B temporal bins and accumulate events into a voxel grid:

$$E(b, x, y) = \sum_{k=1}^{N} \delta(x_k = x, y_k = y)\, \delta(t_k \in \mathcal{B}_b)\, p_k, \tag{6}$$

where $\mathcal{B}_b$ is the b-th temporal bin and $\delta(\cdot)$ denotes the indicator function.

As shown in Fig. 1, the image and event inputs are processed by two parallel GroupMamba-Small encoders. Each encoder consists of a stem layer followed by four hierarchical stages. The stem layer performs local convolutional projection and patch tokenization to produce the stage-0 representations:

$$F_I^{(0)} = \mathrm{Stem}_I(I), \qquad F_E^{(0)} = \mathrm{Stem}_E(E). \tag{7}$$

The modality-specific features are then encoded hierarchically as

$$F_I^{(s)} = \mathcal{G}_I^{(s)}(F_I^{(s-1)}), \qquad F_E^{(s)} = \mathcal{G}_E^{(s)}(F_E^{(s-1)}), \quad s = 1, \ldots, S, \tag{8}$$

where $\mathcal{G}^{(s)}$ denotes the s-th GroupMamba stage and $S = 4$ in our implementation.

Each stage is built from GroupMamba layers. As illustrated in Fig. 3, a GroupMamba layer partitions the input sequence into multiple directional groups and processes them using VSSS blocks, followed by channel modulation and a feed-forward update. This design provides efficient long-range dependency modeling while preserving local structure, making it well suited for dense multimodal segmentation.

3.3 Mamba-Wave Cross-Modal Fusion

The proposed **Mamba-Wave Cross-Modal Fusion (MWCMF)** module is the key component of WaveMamba. Given image and event features at the same scale, MWCMF performs cross-modal interaction through two complementary pathways:

- a **Channel Wave Branch (CWB)**, which emphasizes channel-wise global interaction and temporal-context refinement, and
- a **Spatial Wave Branch (SWB)**, which emphasizes spatially selective interaction and structural refinement.

Given normalized inputs

$$\hat{F}_I = \mathrm{Norm}_I(F_I), \qquad \hat{F}_E = \mathrm{Norm}_E(F_E), \tag{9}$$

MWCMF first computes channel-oriented and spatial-oriented cross-modal features:

$$F_I^c, F_E^c = \mathrm{CCWG}(\hat{F}_I, \hat{F}_E), \tag{10}$$

$$F_I^s, F_E^s = \mathrm{CSWG}(\hat{F}_I, \hat{F}_E). \tag{11}$$

The channel branch outputs are refined by temporal selective state-space modeling,

$$Z_I^c = \mathrm{TemporalSSM}(F_I^c) + \mathrm{Conv}_{1\times1}(F_I^c), \qquad Z_E^c = \mathrm{TemporalSSM}(F_E^c) + \mathrm{Conv}_{1\times1}(F_E^c), \tag{12}$$

while the spatial branch outputs are refined by spatial selective scanning,

$$Z_I^s = \mathrm{SS2D}(F_I^s) + \mathrm{Conv}_{1\times1}(F_I^s), \qquad Z_E^s = \mathrm{SS2D}(F_E^s) + \mathrm{Conv}_{1\times1}(F_E^s). \tag{13}$$

To suppress residual ambiguity after propagation, the channel branch is further modulated by channel refinement gates,

$$\bar{Z}_I^c = Z_I^c \odot \mathrm{CRG}(Z_I^c), \qquad \bar{Z}_E^c = Z_E^c \odot \mathrm{CRG}(Z_E^c), \tag{14}$$

and the spatial branch is modulated by spatial refinement gates,

$$\bar{Z}_I^s = Z_I^s \odot \mathrm{SRG}(Z_I^s), \qquad \bar{Z}_E^s = Z_E^s \odot \mathrm{SRG}(Z_E^s), \tag{15}$$

where $\odot$ denotes element-wise multiplication.

The outputs of the two branches are concatenated and projected back to the original channel dimension:

$$\Delta F_I = \mathrm{Proj}_I\left([\bar{Z}_I^c; \bar{Z}_I^s]\right), \qquad \Delta F_E = \mathrm{Proj}_E\left([\bar{Z}_E^c; \bar{Z}_E^s]\right), \tag{16}$$

followed by residual fusion:

$$\tilde{F}_I = F_I + \mathrm{DropPath}(\Delta F_I), \qquad \tilde{F}_E = F_E + \mathrm{DropPath}(\Delta F_E). \tag{17}$$

Overall, MWCMF allows image and event features to interact through complementary channel-wise and spatial-wise paths, while preserving modality-specific information through residual refinement.

Channel Wave Branch: The CWB models global cross-modal interaction in a channel-aware manner. As shown in Fig. 2(b), it consists of **Cross-Modal Channel-Wave Gating (CCWG)** followed by temporal selective state-space refinement. This branch is intended to capture global channel dependencies while improving temporal consistency, which is particularly useful for event features that contain strong motion-related cues but limited appearance information.

The central idea is to first propagate each modality through a spectral wave operator and then use the propagated response from one modality to gate the other. Let

$$W_I = \mathrm{SWP}(F_I), \qquad W_E = \mathrm{SWP}(F_E), \tag{18}$$

where SWP denotes the Spectral Wave Propagator. CCWG then computes cross-modal channel gates as

$$G_I^c = \sigma\left(\mathrm{Conv}_{1\times1}\big(\mathrm{GAP}(W_E)\big)\right), \tag{19}$$

$$G_E^c = \sigma\left(\mathrm{Conv}_{1\times1}\big(\mathrm{GAP}(W_I)\big)\right), \tag{20}$$

where GAP denotes global average pooling and σ is the sigmoid function. The resulting channel-modulated features are

$$F_I^c = F_I \odot G_I^c, \qquad F_E^c = F_E \odot G_E^c. \tag{21}$$

These features are subsequently refined by a Temporal SSM block and a residual 1×1 convolution.

Spatial Wave Branch: The SWB complements the channel branch by emphasizing spatially selective cross-modal interaction. As shown in Fig. 2(c), it is built on **Cross-Modal Spatial-Wave Gating (CSWG)** followed by SS2D-based refinement.

As in CCWG, the two modalities are first propagated by the Spectral Wave Propagator:

$$W_I = \mathrm{SWP}(F_I), \qquad W_E = \mathrm{SWP}(F_E). \tag{22}$$

CSWG then generates dense spatial gates from the propagated cross-modal responses:

$$G_I^s = \sigma\left(\mathrm{Conv}_{1\times1}(W_E)\right), \tag{23}$$

$$G_E^s = \sigma\left(\mathrm{Conv}_{1\times1}(W_I)\right). \tag{24}$$

The gated outputs are

$$F_I^s = F_I \odot G_I^s, \qquad F_E^s = F_E \odot G_E^s. \tag{25}$$

The resulting features are refined by SS2D blocks, which scan the feature map along multiple directions to capture long-range spatial dependencies.

Spectral Wave Propagator: The **Spectral Wave Propagator (SWP)**, illustrated in Fig. 2(d), is the core operator used in MWCMF. It first performs local encoding with depthwise convolution and linear channel expansion, and then maps the features into the spectral domain using a 2D discrete cosine transform (DCT). We use DCT because it provides an efficient real-valued frequency decomposition for dense spatial feature maps.

Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, SWP first computes

$$U = \mathrm{Linear}\left(\mathrm{DWConv}(X)\right), \tag{26}$$

and splits the result into two parts,

$$[U_0, Z] = \mathrm{Split}(U), \tag{27}$$

where U_0 is used for spectral propagation and Z serves as a gating branch. The spectral state is initialized as

$$\mathcal{U}_0 = \mathrm{DCT}(U_0), \qquad \mathcal{V}_0 = \mathrm{DCT}(U_0). \tag{28}$$

Let T denote a learned frequency embedding. SWP models spectral evolution as

$$\mathcal{Y} = \cos(cT) \odot \mathcal{U}_0 + \frac{\sin(cT)}{c+\epsilon} \odot \left(\mathcal{V}_0 + \frac{\alpha}{2}\mathcal{U}_0\right), \tag{29}$$

where c and α are learnable parameters controlling propagation speed and damping, respectively, and ϵ is a small constant for numerical stability.

This parameterization is inspired by damped wave dynamics in the spectral domain. The cosine term models oscillatory refinement of frequency components, while the damping-related term helps control over-smoothing and preserve modality-specific structure. This is useful in event-image interaction because one modality is sparse and motion-sensitive, whereas the other is dense and appearance-rich.

The propagated response is transformed back to the spatial domain by

$$Y = \mathrm{IDCT}(\mathcal{Y}). \tag{30}$$

Finally, SWP applies normalization and gated output projection:

$$\mathrm{SWP}(X) = \mathrm{Linear}_{out}\Big(\mathrm{LN}(Y) \odot \mathrm{SiLU}(Z)\Big). \tag{31}$$

3.4 Decoder and segmentation prediction

At each scale, MWCMF outputs modality-enhanced image and event features, which are concatenated and projected to form the fused multiscale representation:

$$F^{(s)} = \mathrm{Conv}_{1\times1}\left([\tilde{F}_I^{(s)}; \tilde{F}_E^{(s)}]\right). \tag{32}$$

The set of fused features $\{F^{(1)}, \ldots, F^{(S)}\}$ is then fed into a lightweight SegFormer-style MLP decoder:

$$\hat{Y} = \mathrm{Decoder}\left(F^{(1)}, \ldots, F^{(S)}\right), \tag{33}$$

followed by bilinear upsampling to the original resolution:

$$\hat{M} = \mathrm{Upsample}(\hat{Y}) \in \mathbb{R}^{K\times H\times W}, \tag{34}$$

where K denotes the number of semantic classes.

Table 1: Comparison with RGB and Event based semantic segmentation methods. WaveMamba achieves the best overall mIoU on both benchmarks while maintaining strong pixel accuracy.

Method	Modality	Backbone	Event Rep.	DDD17 mIoU ↑	DDD17 Acc. ↑	DSEC mIoU ↑	DSEC Acc. ↑
SegFormer [3]	RGB	MiT-B2	—	71.05	95.73	71.99	94.97
SegNeXt [4]	RGB	MSCAN-B	—	71.46	95.97	71.55	94.89
EV-SegNet [9]	Event	Xception	6-Channel Image	54.81	89.76	51.76	88.61
ESS [10]	Event	E2ViD	Voxel Grid	61.37	91.08	51.57	89.25
EDCNet-S2D [30]	RGB-Event	ResNet-101	Voxel Grid	61.99	93.80	56.75	92.39
CMX [11]	RGB-Event	MiT-B2	Voxel Grid	67.47	94.20	65.29	92.61
CMNeXt [12]	RGB-Event	MiT-B2	Voxel Grid	66.99	93.82	67.20	93.13
HALSIE [20]	RGB-Event	FCN	Voxel Grid	60.66	92.50	52.43	89.01
CMESS [31]	RGB-Event	E2ViD	Voxel Grid	64.30	92.07	59.53	91.11
OpenESS [32]	RGB-Event	E2VID	Voxel Grid	63.00	91.05	57.21	90.21
SpikingEDN [21]	RGB-Event	FCN	Voxel Grid	72.57	—	58.32	—
SE-Adapter [22]	RGB-Event	SAM	MSP	69.06	95.32	69.77	93.58
EISNet [13]	RGB-Event	MiT-B2	AET	75.03	96.04	73.07	95.12
Spike-BRGNet [33]	RGB-Event	MiT-B2	Voxel Grid	54.72	—	54.95	—
Hybrid-Seg [23]	RGB-Event	FCN	Voxel Grid	67.31	95.07	66.57	94.27
MambaSeg [18]	RGB-Event	Mamba	Voxel Grid	77.56	96.33	75.00	95.71
WaveMamba (Ours)	RGB-Event	Mamba	Voxel Grid	**78.75**	**96.50**	**76.00**	**96.00**

4 Experiments

4.1 Implementation Details

All models are implemented in PyTorch 2.0 and trained on a single NVIDIA RTX-3090 GPU using the AdamW optimizer and pixel-wise cross-entropy loss. Unless otherwise noted, all models are trained for 60 epochs. For **DDD17**, the initial learning rate is set to 2×10^{-4} with a batch size of 8. Events are converted into voxel grids with 10 temporal bins over fixed 50 ms intervals. The input resolution is 200×346. For **DSEC**, the initial learning rate is set to 6×10^{-5} with a batch size of 2. Event streams are represented as voxel grids with 10 bins, where each bin contains 100k events. The input resolution is 440×640. For both datasets, we use standard data augmentation including random horizontal flipping, random cropping, and photometric distortion. Unless otherwise stated, all compared fusion variants share the same lightweight SegFormer-style MLP decoder for controlled comparison. Reproduced baselines are trained under the same data preprocessing, decoder, augmentation, and optimization settings as WaveMamba unless explicitly stated otherwise. To study robustness under degraded visual conditions, we further construct a corrupted DSEC setting by introducing rain and fog effects into approximately 50% of the training images and 50% of the validation images. All methods compared in this setting are retrained under the same protocol.

4.2 Quantitative Comparison with State of the Art

We compare WaveMamba with representative RGB-only, event-only, and RGB-event segmentation methods on DDD17 and DSEC. The results are reported in Table 1. On **DDD17**, WaveMamba achieves **78.75** mIoU and **96.50**% pixel accuracy, outperforming all compared methods, including the MambaSeg baseline. On **DSEC**, WaveMamba reaches **76.00** mIoU and **96.00**% pixel accuracy, again yielding the strongest overall performance among the listed methods. These gains suggest that the proposed wave-guided fusion design improves how dense image semantics and sparse event dynamics complement each other. In particular, compared with transformer-based RGB-event fusion models such as CMX, CMNeXt, and EISNet, WaveMamba achieves higher segmentation accuracy while maintaining a relatively lightweight interaction mechanism.

Table 2 contrasts the complexity of models on DDD17. WaveMamba has the best mIoU and has 94.31M parameters and 34.10G MACs. The proposed model contain low parameters when compared to MambaSeg with the accuracy gain being 35.5 MACs with a 97.55M parameter count. WaveMamba offers a good accuracy-compute trade-off, compared with transformer-based approaches, such as SegFormer, SwiftFormer, CMX, and CMNeXt. These findings suggest that multimodal fusion with wave guidance can enhance the quality of semantic segmentation.

Table 2: Model complexity comparison on DDD17.

Method	Backbone	Params (M)	MACs (G)	mIoU (%)
EV-SegNet	CNN	29.09	73.62	54.81
EDCNet	CNN	23.06	6.14	61.99
SegFormer	Transformer	51.54	19.78	71.05
SwiftFormer	Transformer	64.61	23.88	74.31
CMX	Transformer	66.56	16.29	67.47
CMNeXt	Transformer	58.68	16.32	66.99
EISNet	Transformer	34.39	17.30	75.03
MambaSeg	Mamba	97.55	35.59	77.56
WaveMamba (Ours)	Mamba	94.31	34.10	**78.75**

Table 3: Ablation study of WaveMamba on DDD17 and DSEC.

	Variant	CCWG	CSWG	T-SSM	SS2D	CRG	SRG	DDD17 mIoU	DSEC mIoU
Backbone	MambaSeg† (VMamba + DDIM)							77.23	—
	GroupMamba + DDIM *(our baseline)*							77.80	74.70
Fusion	+ CCWG only	✓						77.92	75.41
	+ CSWG only		✓					78.11	75.58
	CCWG+CSWG (no SSM)	✓	✓					78.33	75.70
	+ Temporal-SSM (channel)	✓	✓	✓				78.52	75.85
	+ SS2D (spatial)	✓	✓		✓			78.44	75.78
	− CRG	✓	✓	✓	✓		✓	78.16	75.62
	− SRG	✓	✓	✓	✓	✓		78.30	75.71
	Fast wave variant	✓	✓	✓	✓	✓	✓	78.21	75.20
	Full MWCMF(WaveMamba)	✓	✓	✓	✓	✓	✓	**78.75**	**76.00**

4.3 Ablation Study

We conduct a systematic ablation study in two parts: (1) isolating the contribution of the backbone relative to MambaSeg, and (2) progressively evaluating the individual components of MWCMF while keeping the GroupMamba encoder fixed. Replacing VMamba-T with GroupMamba-Small while retaining DDIM fusion improves DDD17 performance from 77.23 to 77.80 mIoU. This result indicates that directional multi-group scanning provides a stronger representational foundation even before changing the fusion design. Starting from the GroupMamba + DDIM baseline (77.80 / 74.70 on DDD17 / DSEC), we progressively introduce the components of MWCMF. Using only CCWG or only CSWG already improves performance, with CSWG producing the larger gain. This suggests that spatial correspondence is a particularly important bottleneck in event-image interaction. Combining the two wave-gating mechanisms yields a further improvement, indicating that the channel-wise and spatial-wise pathways are complementary.

Adding Temporal-SSM and SS2D each provides additional gains. Temporal-SSM contributes slightly more on DDD17, where events are binned over fixed 50 ms intervals and temporal regularity is stronger, whereas SS2D contributes slightly more on DSEC, where the event representation is more spatially distributed. The full MWCMF delivers the highest performance on both benchmarks, with the highest performance of 78.75 mIoU on DDD17 and 76.00 mIoU on DSEC. The removal of channel residual gate or spatial residual gate leads to performance decreases indicating that post-propagation refinement is still significant to eliminate the residual cross-modal ambiguity. Replacement of the learnable Spectral Wave Propagator with a fixed fast-wave variant also harms performance, which is an indicator that adaptive spectral modulation is a significant component of the final design.

On the whole, this ablation study confirms three observations: **(1)** GroupMamba enhances the representational backbone as compared to the MambaSeg and EISNet baselines; (2) spatial wave gating contributes slightly more than channel wave gating; (4) both channel wave branch and spatial wave branch are required in both instances, in order to achieve the best performance.

Qualitative Results Figures 4 and 5 present qualitative comparisons on DDD17 and DSEC, respectively. Each figure shows the event input, image input, predictions from representative baselines, the WaveMamba prediction, and the ground-truth labels. On **DDD17**, WaveMamba produces cleaner object boundaries and more complete predictions for small or fast-moving objects such as distant

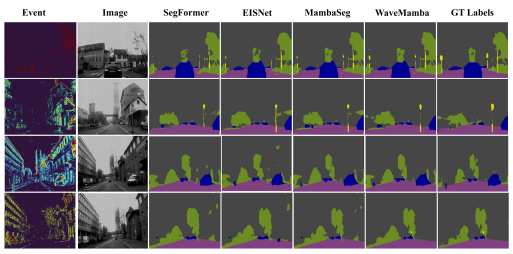

Figure 4: Qualitative comparison on the DDD17 validation set.

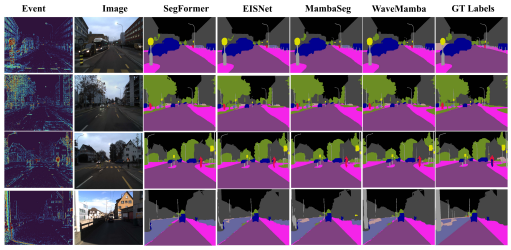

Figure 5: Qualitative comparison on the DSEC validation set.

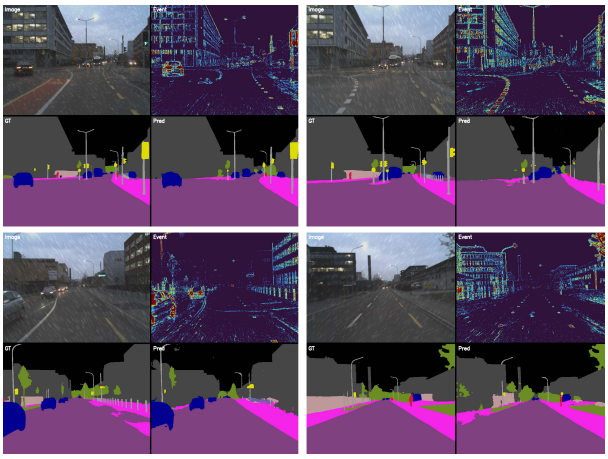

Figure 6: Robustness under rain and fog corruption on DSEC.

vehicles and cyclists. Compared with MambaSeg and EISNet, it better preserves thin structures and reduces fragmented predictions near motion boundaries. On **DSEC**, WaveMamba maintains more consistent segmentation in high-resolution urban scenes. In particular, it yields stronger predictions for roads, sidewalks, poles, traffic signs, and vehicles, while reducing false positives in visually ambiguous regions such as shadows, reflective surfaces, and low-contrast areas.

Robustness Under Adverse Weather We further evaluate robustness under degraded visual conditions in Fig. 6. In this experiment, the DSEC training and validation data are partially augmented with rain and fog corruptions. WaveMamba achieves **74.0** mIoU under the corrupted setting, compared with **63.0** mIoU and **61.1** mIoU for the reproduced MambaSeg and EISNet baselines. As shown in Fig. 6, WaveMamba reliably segments distant cars, traffic lights, and road boundaries even when the RGB image is visibly degraded by fog or rain.

5 Further Analysis

Attributing the fusion gains. Replacing VMamba-T by GroupMamba-S while retaining DDIM raises DDD17 mIoU from 77.23 to 77.80 ($+0.57$). With the GroupMamba backbone fixed, dual-path gating adds $+0.53/+1.00$ on DDD17/DSEC, rising to $+0.95/+1.30$ with full MWCMF (Table 4). With identical gating and refinement, DCT-based SWP outperforms the FFT radial filter and the more flexible Fourier CWFF, which additionally learns a complex frequency filter and cross-modal coupling. This supports the tested operator, not a physical wave-equation claim.

Consistency across configurations. For MambaSeg and WaveMamba we evaluated three training configurations, varying the initial learning rate and schedule while keeping the benchmark protocol consistent. WaveMamba obtains 78.69 ± 0.08 versus 77.21 ± 0.02 mIoU on DDD17 and 75.92 ± 0.08 versus 74.66 ± 0.01 on DSEC. Their observed performance ranges do not overlap. These vary

Table 4: Fusion/operator ablations (mIoU, %); last three rows differ only by spectral operator.

Configuration	DDD17	DSEC
GroupMamba + DDIM	77.80	74.70
+ CCWG + CSWG (without SSM refinement)	78.33	75.70
MWCMF with FFT radial filter	78.21	75.20
MWCMF with Fourier CWFF	78.50	75.40
MWCMF with DCT-based SWP (ours)	**78.75**	**76.00**

optimization configurations, not random seeds. MambaSeg and EISNet were reproduced under a common pipeline; other Table 1 baselines are publication-reported.

Robustness under matched corruption protocols. We compare *unseen* rain/fog (clean training) with *corruption-aware* training (rain/fog augmentation). Under matched protocols (Table 5), WaveMamba achieves 65.4 mIoU, exceeding MambaSeg by 2.4 and EISNet by 4.3 points, with the smallest relative degradation from clean performance. With corruption-aware training, it attains 74.0 mIoU, but does not have the smallest relative degradation. Reducing DSEC event bins from 10 to 1 further reduces our mIoU from 76.00 to 74.50.

Table 5: Matched DSEC rain/fog evaluation (mIoU, %); unseen vs. corruption-aware training.

Method	Clean	Unseen	Rel. drop	Aware
EISNet	73.07	61.1	16.4%	71.5
MambaSeg	75.00	63.0	16.0%	72.9
WaveMamba	**76.00**	**65.4**	**13.9%**	**74.0**

Measured inference efficiency. At DDD17 200×346, batch 1, FP32, WaveMamba uses 94.31M parameters (26.44M in fusion), versus 97.55M (33.54M in fusion) for MambaSeg; measured memory is 0.494 versus 0.522 GB, and throughput is 75 versus 70 FPS on RTX 3090 and 44 versus 41 FPS on RTX 4060 Ti. These improvements apply *within dual-branch Mamba models*; the network is not lightweight in absolute terms.

6 Limitations

The DCT frequency embedding is resolution-dependent, FFT variants are asymptotically cheaper. Plain non-spectral gating, temporal asynchrony, event noise, modality dropout, calibration drift, cross-dataset transfer, and on-device latency remain untested. The comparisons above also vary optimizer configurations rather than random seeds.

7 Conclusion

We presented **WaveMamba**, a dual-branch event-image semantic segmentation framework that combines parallel GroupMamba encoders with the proposed **Mamba-Wave Cross-Modal Fusion (MWCMF)** module. The proposed **MWCMF** module performs implicit cross-modal propagation through channel-wise and spatial-wise gating in the spectral domain, followed by Mamba-based refinement. Extensive experiments on the DDD17 and DSEC benchmarks demonstrate that WaveMamba achieves state-of-the-art performance, reaching **78.75** mIoU on DDD17 and **76.00** mIoU on DSEC. In addition, WaveMamba exhibits strong robustness under adverse weather conditions. When trained and evaluated on rain- and fog-augmented DSEC data, it attains **74.0** mIoU, substantially outperforming the RGB-event baselines. Ablation studies further verify that each component of MWCMF contributes positively, with the full model providing the best balance of accuracy, robustness, and efficiency.

Acknowledgment

This work was supported by the Guangdong Provincial Key Laboratory of Fully Actuated System Control Theory and Technology, Southern University of Science and Technology, Shenzhen 518055, China. The authors' work was also supported in part by the Guangdong Science and Technology Program under Grant No. 2024B1212010002, and the National Natural Science Foundation of China (NSFC) under Project Grant No. W2533171 and 62373183.

References

[1] Zhao, W., Zhang, R., Wang, Q., Cheng, G. and Huang, K., 2025. Bfanet: Revisiting 3d semantic segmentation with boundary feature analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 29395-29405).

[2] Yin, B.W., Cao, J.L., Cheng, M.M. and Hou, Q., 2025. Dformerv2: Geometry self-attention for rgbd semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19345-19355).

[3] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

[4] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu. SegNeXt: Rethinking convolutional attention design for semantic segmentation. In *NeurIPS*, 2022.

[5] Long, X., Zhu, X., Guo, F., Zhang, W., Gu, Q., Chen, C. and Gu, F., 2025, October. SLTNet: Efficient Event-based Semantic Segmentation with Spike-driven Lightweight Transformer-based Networks. In 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 4331-4338). IEEE.

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

[7] Zheng, X., Liu, Y., Lu, Y., Hua, T., Pan, T., Zhang, W., Tao, D. and Wang, L., 2023. Deep learning for event-based vision: A comprehensive survey and benchmarks. arXiv preprint arXiv:2302.08890.

[8] Shiba, S., Klose, Y., Aoki, Y. and Gallego, G., 2024. Secrets of event-based optical flow, depth and ego-motion estimation by contrast maximization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), pp.7742-7759.

[9] I. Alonso and A. C. Murillo. EV-SegNet: Semantic segmentation for event-based cameras. In *CVPR Workshops*, 2019.

[10] Z. Sun, N. Messikommer, D. Gehrig, and D. Scaramuzza. ESS: Learning event-based semantic segmentation from still images. In *ECCV*, 2022.

[11] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE TITS*, 2023.

[12] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiss, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *CVPR*, 2023.

[13] B. Xie, Y. Deng, Z. Shao, and Y. Li. EISNet: A multi-modal fusion network for semantic segmentation with events and images. *IEEE TMM*, 2024.

[14] Yao, Z. and Chuah, M.C., 2025, February. Event-guided low-light video semantic segmentation. In 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 3330-3341). IEEE.

[15] Gu, A., Goel, K. and Ré, C., 2021. Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396.

[16] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[17] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu. VMamba: Visual state space model. In *NeurIPS*, 2024.

[18] F. Gu, Y. Li, X. Long, K. Ji, C. Chen, Q. Gu, and Z. Ni. MambaSeg: Harnessing Mamba for accurate and efficient image-event semantic segmentation. In *AAAI*, 2026.

[19] Zhao, Y., Lyu, G., Li, K., Wang, Z., Chen, H., Yang, Z. and Deng, Y., 2025, April. ESEG: event-based segmentation boosted by explicit edge-semantic guidance. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 10, pp. 10510-10518).

[20] S. Das Biswas, A. Kosta, C. Liyanagedera, M. Apolinario, and K. Roy. HALSIE: Hybrid approach to learning segmentation by simultaneously exploiting image and event modalities. In *WACV*, 2024.

[21] R. Zhang, L. Leng, K. Che, H. Zhang, J. Cheng, Q. Guo, J. Liao, and R. Cheng. Accurate and efficient event-based semantic segmentation using adaptive spiking encoder-decoder network. *IEEE TNNLS*, 2024.

[22] B. Yao, Y. Deng, Y. Liu, H. Chen, Y. Li, and Z. Yang. SAM-Event-Adapter: Adapting segment anything model for event-RGB semantic segmentation. In *ICRA*, 2024.

[23] H. Li, Y. Peng, J. Yuan, P. Wu, J. Wang, Y. Zhang, and X. Sun. Efficient event-based semantic segmentation via exploiting frame-event fusion: A hybrid neural network approach. In *AAAI*, 2025.

[24] Chakravarthi, B., Verma, A.A., Daniilidis, K., Fermuller, C. and Yang, Y., 2024, September. Recent event camera innovations: A survey. In European conference on computer vision (pp. 342-376). Cham: Springer Nature Switzerland.

[25] Messikommer, N., Fang, C., Gehrig, M. and Scaramuzza, D., 2023. Data-driven feature tracking for event cameras. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5642-5651).

[26] Sundar, V., Dutson, M., Ardelean, A., Bruschini, C., Charbon, E. and Gupta, M., 2024. Generalized event cameras. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 25007-25017).

[27] Gehrig, M., Muglikar, M. and Scaramuzza, D., 2024. Dense continuous-time optical flow from event cameras. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(7), pp.4736-4746.

[28] J. Ruan, J. Li, and S. Xiang. VM-UNet: Vision Mamba UNet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.

[29] Z. Wang, C. Li, H. Xu, and X. Zhu. Mamba YOLO: SSMs-based YOLO for object detection. *arXiv preprint arXiv:2406.05835*, 2024.

[30] Zhang, J., Yang, K. and Stiefelhagen, R., 2021. Exploring event-driven dynamic context for accident scene segmentation. IEEE Transactions on Intelligent Transportation Systems, 23(3), pp.2606-2622.

[31] Xie, C., Gao, W. and Guo, R., 2024. Cross-modal learning for event-based semantic segmentation via attention soft alignment. IEEE Robotics and Automation Letters, 9(3), pp.2359-2366.

[32] Kong, L., Liu, Y., Ng, L.X., Cottereau, B.R. and Ooi, W.T., 2024. 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).

[33] Long, X., Zhu, X., Guo, F., Chen, C., Zhu, X., Gu, F., Yuan, S. and Zhang, C., 2024. 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), pp.2712-2724.

8 Technical appendices and supplementary material

8.1 Comparison of Fusion Modules

To isolate the effect of the proposed fusion design, we compare MWCMF with several representative fusion modules under the same backbone and decoder setting. As shown in Table 6, MWCMF achieves the best performance on DDD17, reaching **78.75** mIoU and **96.50**% pixel accuracy. MWCMF outperforms FFM, MRFM, the EISNet fusion module, and DDIM from MambaSeg under the same controlled setting. This result supports the view that wave-guided cross-modal propagation provides a stronger interaction mechanism than conventional rectification- or attention-style fusion in this architecture.

Table 6: Comparison of fusion modules on DDD17 under a fixed backbone and decoder.

Fusion Module	mIoU (%)	Acc. (%)
CSF [23]	76.85	96.24
FFM [11]	76.77	96.16
MRFM [13]	76.58	95.97
DDIM [18]	77.80	96.23
MWCMF (Ours)	**78.75**	**96.50**

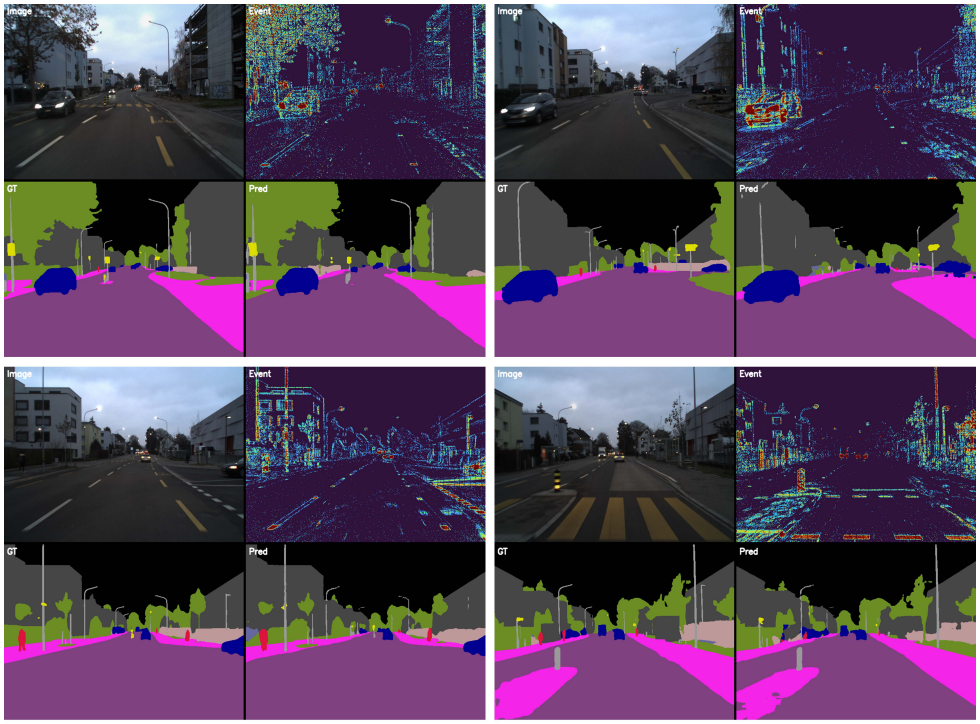

Figure 7: Additional qualitative results on DSEC. Each example shows the event input, image frame, ground-truth segmentation, and the prediction produced by WaveMamba. Across these sequences, WaveMamba preserves road layout, thin structures, and vehicle boundaries while reducing fragmented regions and false positives.

8.2 Additional Visualization Results on DSEC

We present extended visualization results on both DSEC and DDD17 to further illustrate the behavior of the proposed cross-modal fusion mechanism under diverse scene conditions. For each example, we show the event representation, the corresponding image frame, the ground-truth segmentation mask, and the prediction produced by WaveMamba. These examples complement the qualitative comparisons in the main paper and provide a more detailed view of the model's behavior across different sequences and lighting conditions. Figure 7 presents four additional qualitative examples from DSEC sequences. Each example includes the event input, the corresponding image frame, the ground-truth semantic mask, and the segmentation predicted by WaveMamba. These results highlight the model's ability to preserve road structure, recover thin objects such as poles and traffic signs, and maintain consistent predictions for vehicles and sidewalks in cluttered urban scenes. The examples also show that the proposed fusion design remains effective when event responses are sparse or spatially uneven.

8.3 Additional Visualization Results on DDD17

Figure 8 shows additional qualitative examples on DDD17. As in the DSEC results, each row includes the event input, image frame, ground-truth mask, and WaveMamba prediction. These examples further demonstrate that the model produces cleaner semantic regions and sharper object boundaries, especially for small or distant structures. The results also indicate that WaveMamba can better preserve scene layout in challenging motion-dominant scenarios.

8.4 DSEC Visualization Under Dark-Time Conditions

Figure 9 presents additional qualitative results on DSEC during dark-time conditions. These examples are included to highlight the behavior of WaveMamba when image appearance is degraded and the event modality becomes particularly informative. Each example contains the event input, image frame, ground-truth segmentation, and the corresponding WaveMamba prediction. Even in low-visibility scenes, the model preserves the main road structure and recovers important semantic categories such as vehicles, poles, and traffic signs with relatively stable boundaries.

Table 7: Ablation study of the event bin number N on DSEC. Increasing the number of temporal bins improves performance, with the best result achieved at $N = 10$.

Method	Event Bin N	mIoU ↑	Accuracy ↑
WaveMamba	1	74.50	95.6
WaveMamba	3	74.90	95.6
WaveMamba	5	75.47	95.7
WaveMamba	10	**76.00**	**96.0**

8.5 Ablation on the Number of Event Bins

We further study the effect of the event bin number N on DSEC. Table 7 reports the performance of WaveMamba with different temporal bin settings. As N increases from 1 to 10, segmentation performance improves consistently, indicating that a richer temporal discretization provides more informative event representations for dense prediction. The best result is obtained with $N = 10$, which achieves **76.00** mIoU and **96.0**% accuracy.

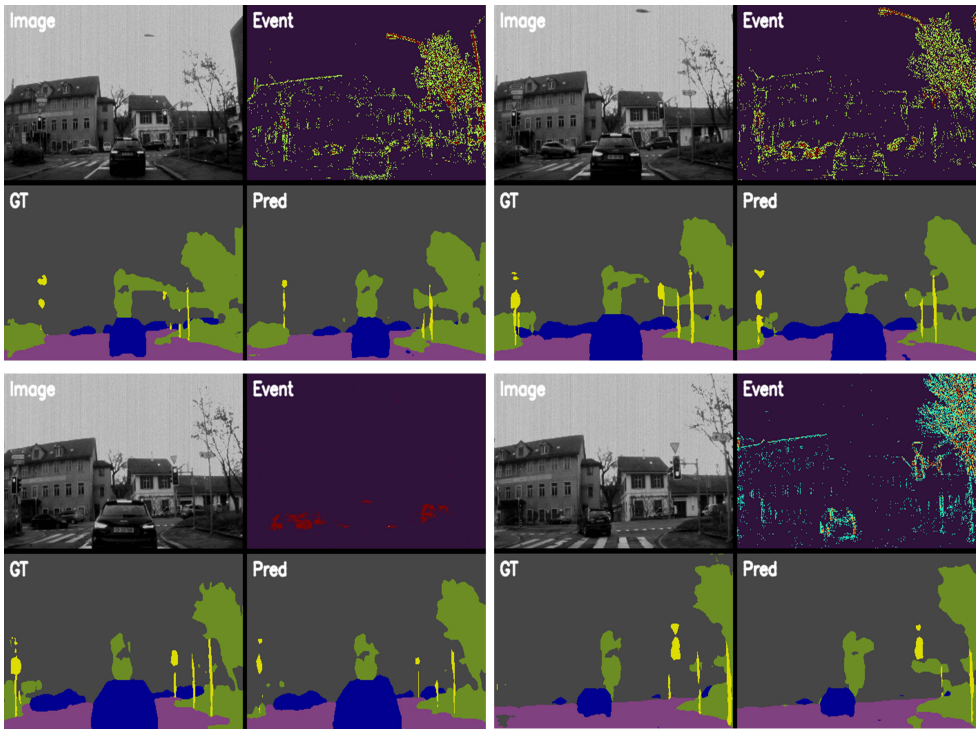

Figure 8: Additional qualitative results on DDD17. Each row shows the event input, image frame, ground truth, and WaveMamba prediction. The model produces cleaner object boundaries and more complete semantic regions, particularly for small or distant objects in motion-heavy scenes.

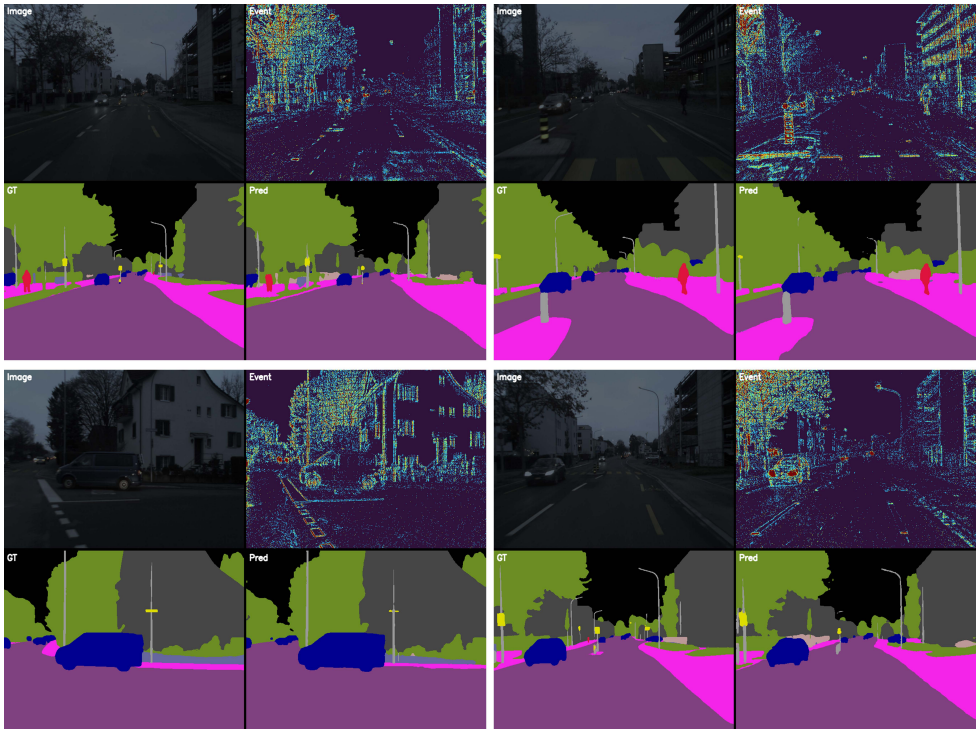

Figure 9: Additional qualitative results on DSEC during dark-time conditions. Each example shows the event input, image frame, ground truth, and WaveMamba prediction. Even when the image modality is visually degraded, WaveMamba maintains consistent road structure and recovers important scene elements such as vehicles, poles, and traffic signs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the proposed WaveMamba architecture, the MWCMF fusion module, and the main empirical claims on DDD17, DSEC, and corrupted DSEC. The claims are consistent with the reported experimental results in the main paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The current manuscript include an short section where Robustness Under Adverse Weather limitations is discussed. We evaluate robustness under degraded visual conditions by constructing a corrupted DSEC setting with rain and fog augmentations applied to a portion of the training and validation data. The corruption is applied to approximately 50% of images, meaning the model is not evaluated under fully corrupted conditions across all sequences. Furthermore, As shown in the ablation study in Appendix, WaveMamba's performance on DSEC degrades noticeably when the number of temporal event bins is reduced. With $N{=}1$ bin, mIoU drops to 74.50 compared to 76.00 at $N{=}10$, indicating that the model benefits from richer temporal discretization. This suggests that in scenarios where only sparse or low-frequency event data is available, further architectural adaptation may be required to maintain robust segmentation quality. We leave the development of bin-efficient event representations for future work.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper is an empirical method paper and does not present formal theoretical results such as theorems, lemmas, or proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The manuscript specifies the model architecture, datasets, event representation, optimizer, learning rates, batch sizes, training epochs, image resolutions, augmentations, decoder setting, and the corrupted DSEC protocol. These details are sufficient to understand and reproduce the main experimental setup at the level required for the paper's claims.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplementary material includes the full implementation code, training and evaluation scripts, pretrained model checkpoints for both DDD17 and DSEC-Semantic, and a results video demonstrating qualitative performance under standard, rain, fog, and dark-time conditions. The datasets used are publicly available benchmarks. Code and checkpoints are also accessible at `https://github.com/adeelferozmirza/WaveMamba`.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental section specifies optimizer, epochs, learning rates, batch sizes, event binning, image resolutions, augmentation choices, decoder consistency, and the corrupted DSEC evaluation setting. These details are sufficient to understand how the reported results were obtained.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports point estimates only and does not include standard deviations, confidence intervals, or repeated-run statistics. No formal significance analysis is currently provided.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper specifies the GPU type used for training but does not report memory configuration, runtime per experiment, total GPU hours, or total project compute. Therefore the compute disclosure is only partial.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, this work conforms to the NeurIPS Code of Ethics. The paper presents a methodological contribution evaluated on existing benchmarks and does not involve human-subject experiments or the release of high-risk generative models.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper does not currently include a dedicated broader-impact discussion. Although the work is relevant to autonomous driving and robotics, the manuscript does not explicitly discuss possible benefits, risks, or mitigation strategies.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk generative model, scraped dataset, or similar asset with elevated misuse risk. The work focuses on a segmentation architecture evaluated on existing benchmarks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper cites prior methods and benchmark datasets, but it does not explicitly document the licenses or terms of use of the datasets, codebases, or reproduced baselines used in the experiments.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The supplementary material provides the WaveMamba codebase with inline documentation, configuration files for both benchmarks, pretrained checkpoints for DDD17 and DSEC-Semantic, and a results video illustrating segmentation predictions under standard and degraded visual conditions. Installation instructions and dataset preparation guidelines are included in the README at `https://github.com/adeelferozmirza/WaveMamba`.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects, so IRB approval and participant-risk disclosure are not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the proposed model, methodology, or experimental pipeline. Any use of language models for writing or editing would not constitute a core scientific component of the research.