Towards Robust Event-Based Depth Estimation: Bridging Synthetic and Real Domains with Motion Adaptation

Yuzhe Ji[1], **Haotian Wang**[1], **Yijie Chen**[1], **Xiang Cheng**[3], **Liuqing Yang**[1, 2], **Xinhu Zheng** [1, 2 †]

[1] Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2] Hong Kong University of Science and Technology, Hong Kong SAR, China
[3] Peking University, Beijing, China
{yji755, ychen324}@connect.hkust-gz.edu.cn, xiangcheng@pku.edu.cn,
{haotianwang, lqyang, xinhuzheng}@hkust-gz.edu.cn

Abstract

Event cameras offer microsecond latency and high dynamic range, making them particularly suitable for safety-critical 3D perception in autonomous driving scenarios with challenging lighting conditions. Yet existing methods often struggle to generalize to out-of-domain environments due to the limited availability of diverse training data. While synthetic data offers an easily accessible alternative, it introduces a significant sim-to-real gap, particularly in motion patterns. We tackle this challenge by introducing **M**otion-**A**daptation Mamba (MA-Mamba), a dual-track framework that advances both architecture and data augmentation. At the architectural level, we introduce a lightweight Spatial-Temporal Association module that captures motion-induced appearance variations at arbitrary scales, and an Adaptive Memory Balancing module, built on the Mamba state-space framework, that adaptively filters memory updates to maintain stable scene context under diverse dynamics. At the data level, we design event-oriented augmentations that simulate varied motion patterns and apply priority-based masked sequence modeling to strengthen long-range spatio-temporal reasoning. Trained solely on synthetic data, MA-Mamba delivers substantial zero-shot gains on multiple real-world benchmarks, demonstrating strong robustness and generalizability.

Introduction

Achieving reliable 3D perception with vision sensors alone remains a central goal for both academia and industry. Recent advances enable accurate depth estimation from RGB images, allowing vision-only pipelines to approach LiDAR-based performance across multiple 3D perception tasks (Liu et al. 2024a; Yang et al. 2023; Peng et al. 2023). Nevertheless, these systems still struggle under low-light or high-contrast conditions, such as nighttime driving or tunnel entrances, where RGB cameras fail to capture sufficient detail for robust perception.

Event cameras, on the other hand, record only per-pixel brightness changes asynchronously on a logarithmic scale rather than absolute intensities. This mechanism yields an ultra-high dynamic range ($>$ 120 dB, compared with 60 dB for standard cameras) and maintains reliable performance from moonlight to daylight. Because they transmit only sparse brightness changes, event cameras also reduce power consumption (Gallego et al. 2020). These properties make event-based vision a compelling and lightweight alternative to conventional cameras for resilient 3D perception.

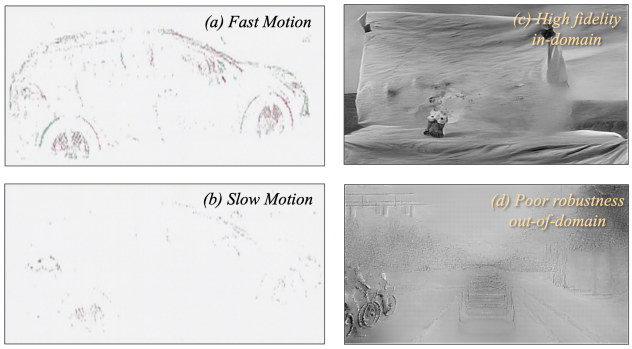

Figure 1: (a–b) Different motions of the same car in a scene trigger distinct event streams (Kugele et al. 2023). (c) The seminal E2VID model achieves high-fidelity reconstructions on familiar motion patterns, while (d) suffering severe degradation under unseen motions.

Nevertheless, event cameras fire asynchronously, producing sparse spatio-temporal event streams that are information-rich yet challenging to process. The most prominent issue limiting deep learning approaches for event cameras is poor robustness. For instance, the seminal E2VID model (Rebecq et al. 2019) can reconstruct impressive HDR videos of complex, subtle dynamics from events, such as bursting balloons or bullets piercing cups. Yet its outputs on simple traffic scenes exhibit severe ghosting and missing details, as illustrated in Fig. 1 (c-d). This poor robustness stems from *motion variability*: each event encodes a binary brightness change driven by optical flow from both ego- and object motion. As shown in Fig. 1 (a-b), changes in an object's motion can markedly alter the data distribution, yielding far greater variability in event streams than in RGB images. Stationary or slow-moving objects may generate few or no events over extended periods, requiring neural net-

†Corresponding authors

works to retain long-term memory and adapt it in response to motion-induced changes. E2VID's authors (Rebecq et al. 2019) also attribute poor generalization to its recurrent components, which fail to track spatial feature changes and thus cannot adapt memory states appropriately.

Existing methods for improving robustness rely on scaling datasets, often requiring millions of labeled or unlabeled images (Ranftl et al. 2020; Yang et al. 2024). This is impractical for event-based vision because real-world event streams are difficult to collect, and dense ground-truth depth for sparse events is costly to obtain. Synthetic data are easier to obtain but typically suffer from unmodeled real-world dynamics, resulting in a substantial sim-to-real gap.

Recent work has largely overlooked the generalization challenge posed by motion variability in event cameras. Asynchronous models such as spiking neural networks or graph neural networks (Cordone, Miramond, and Thierion 2022; Schaefer, Gehrig, and Scaramuzza 2022) are difficult to optimize and thus achieve only moderate in-domain accuracy (Zubic, Gehrig, and Scaramuzza 2024). Frame-based methods aggregate events into short temporal windows and exploit standard vision backbones for stronger optimization. Many recent approaches favor expressive but computationally expensive spatial encoders (Gehrig and Scaramuzza 2023; Liu et al. 2024b) and temporal modules such as bidirectional LSTMs and self-attention mechanisms (Ren et al. 2024; Zhang et al. 2022), which increase model complexity and exacerbate overfitting when training data are limited.

To address these robustness and generalization challenges, we propose a dual strategy integrating targeted model architecture design and event-specific data augmentation. On the model side, we introduce a lightweight Spatial-Temporal Association module that robustly captures motion-induced variations across arbitrary spatial scales with a parameter-invariant design. Furthermore, we devise an Adaptive Memory Balancing module that inherits the long-range dependency modeling capability of state-space models (SSMs) and uses spatial-association cues as adaptive control signals for dynamic memory updates.

On the data side, we propose simple yet effective augmentations specifically tailored to event streams. We employ Random Motion Pattern Generation to simulate acceleration and deceleration patterns within individual scenarios, enriching spatio-temporal variability. Additionally, Priority-based Masked Sequence Modeling compels the temporal module to reconstruct depth information of masked spatial regions, thus significantly strengthening its robustness. Consequently, our model, trained solely on synthetic data, achieves robust zero-shot depth estimation performance when tested on real-world event datasets.

In summary, our contributions are as follows:

- We present a pioneering event-based depth estimation method that robustly adapts to complex real-world motion patterns and bridges the sim-to-real gap.
- Our approach combines a lightweight Spatial-Temporal Association module, an Adaptive Memory Balancing module, and event-oriented augmentations that jointly enhance data efficiency and cross-dataset generalization.
- Trained solely on synthetic data, our method outperforms state-of-the-art methods by up to 59% in zero-shot tests on common real-world datasets MVSEC and DSEC.

Related Work

Event-based Monocular Depth Estimation

Depth is a crucial cue for 3D perception tasks such as robot navigation (Dong et al. 2022) and autonomous driving (Wang et al. 2019). Advances in deep learning now enable dense depth estimation directly from raw event streams. Recent methods have progressively adopted stronger spatial encoders, ranging from deeper CNNs to Transformer blocks, and have upgraded temporal processing by replacing basic LSTMs with attention-based modules to capture long-range context (Hidalgo-Carrió, Gehrig, and Scaramuzza 2020; Gehrig et al. 2021a; Pan, Cao, and Wang 2024; Zhang et al. 2022; Liu et al. 2022). Although such larger networks attain high in-domain accuracy, the limited availability of event data often leads to poor generalization to unseen scenes. Robust depth estimation has been extensively explored in the RGB domain (Ranftl et al. 2020; Hu et al. 2024; Yang et al. 2024), yet the robustness of event-based monocular depth estimation remains largely unexplored.

Event-oriented Data Augmentation

Data augmentation is widely adopted to increase data efficiency. While generic image augmentations, like flipping or scaling, can be applied after converting event streams into frames, they fail to reflect the sparse, noisy, and temporally structured nature of real event data. Consequently, previous studies have proposed several augmentation strategies tailored to event data. EventDrop (Gu et al. 2021) randomly discards events within selected spatial and temporal regions. EventAug (Tian et al. 2024) masks separately computed spatial and temporal areas to mimic object variations but overlooks spatial integrity and temporal continuity. CutMix-style methods such as EventZoom (Duan et al. 2021), EventRPG (Sun et al. 2024), and EventMix (Shen, Zhao, and Zeng 2023) enhance scene diversity and perform well on classification and object-recognition tasks, yet they distort 3D perspective priors and are unsuitable for geometry-aware 3D perception tasks. To overcome these limitations, we introduce augmentation strategies that simulate richer motion patterns and spatio-temporal relations while keeping the 3D priors, thereby enabling more robust depth estimation.

State Space Models

HiPPO (Gu et al. 2020a) equips SSMs with long-range memory by selecting the A matrix. S4 and S4D (Gu, Goel, and Ré 2021; Gu et al. 2022) reduce computational costs through structured-matrix decompositions. Mamba (Gu and Dao 2023) employs input-dependent selective scanning, enabling data-scalable training consistent with scaling laws.

Zubić et al. (Zubic, Gehrig, and Scaramuzza 2024) and Ren et al. (Ren et al. 2025) apply SSMs to event data, replacing conventional recurrent networks to speed up training and enhance long-sequence memory. SMamba (Yang et al. 2025) leverages event-stream sparsity by scanning only

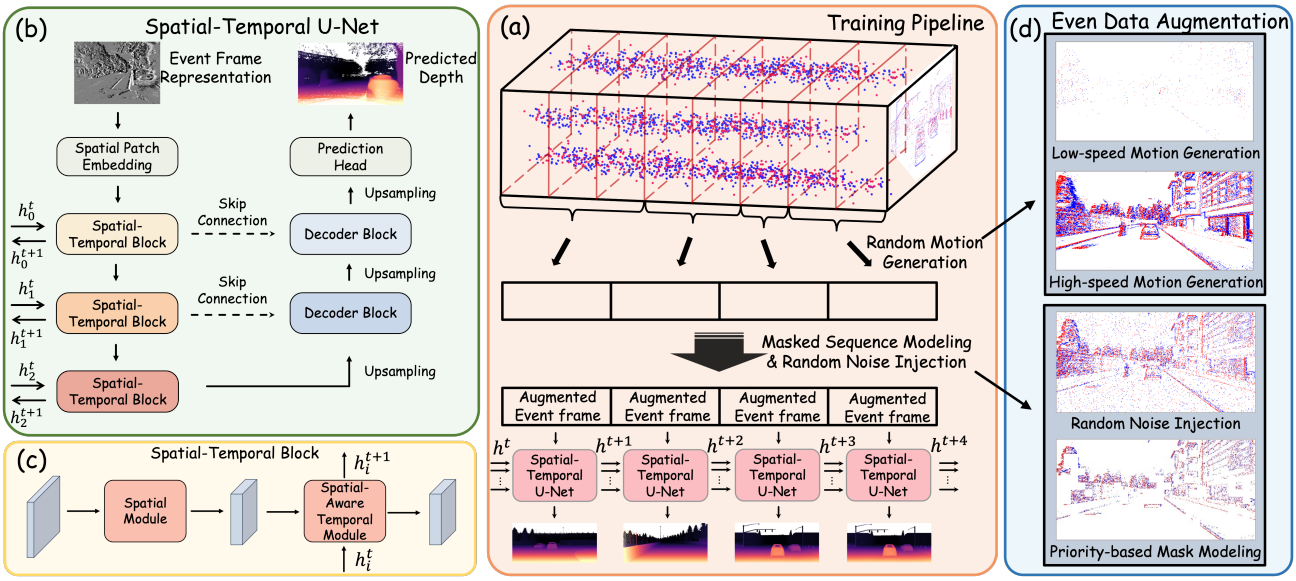

Figure 2: Training pipeline overview (a). An input event frame sequence is first subjected to various event-oriented augmentations depicted in Fig. 2 (d), producing an enriched stream that increases data diversity. (b) The augmented sequence is processed by a recurrent U-Net, which extends a conventional U-Net with temporal feedback at every scale. (c) Within each encoder stage, sparse spatial features are extracted by a spatial module and then fused into temporal memory through our novel Spatial-Aware Temporal Module, which dynamically conditions temporal updates on the current motion-induced spatial context.

information-dense tokens and serves as a fast spatial feature extractor. Prior work still treats spatial and temporal cues in isolation, we instead couple a lightweight spatial-temporal association module with adaptive memory balancing, enabling robust memory updates under diverse motion patterns.

Methodology

This section first formalizes the event-frame representation and its state-space formulation. We then detail the Spatial–Temporal Association (STA) module and the Adaptive Memory Balancing (AMB) module, which couple the spatial indicator to adaptive memory updates. Finally, we introduce two event-oriented data-augmentation techniques: Random Motion Pattern Generation and Priority-based Masked Sequence Modeling. Together, these augmentations and the updated architecture markedly improve the robustness of depth estimation to unseen real-world scenes.

Preliminaries

Event Processing and Representation An event is generated when the logarithmic intensity at pixel (x, y) changes by more than a threshold C within a short interval $(t-\Delta t, t)$. Its polarity $p \in \{-1, 1\}$ records the sign of the change, so each event e_i is fully specified by the four-tuple (p, t, x, y). A stream of events is therefore a list of such tuples.

We adopt the lightweight, empirically validated event-frame representation of (Gehrig and Scaramuzza 2023; Zubic, Gehrig, and Scaramuzza 2024). For the interval $[t_a, t_b)$, we accumulate events by polarity, spatial coordinate, and temporal bin τ to form a four-dimensional tensor $E(p, \tau, x, y)$ as depicted in Eq. 1. The first axis stores the two polarities, the second comprises T temporal slices obtained by uniformly partitioning the interval, and the last two match the sensor resolution $H \times W$.

Flattening the polarity and temporal axes yields a $(2T, H, W)$ tensor that can be processed directly by general vision backbones while offering much finer temporal resolution than its RGB counterpart.

$$E(p,\tau,x,y) = \sum_{e_k \in \mathcal{E}} \delta(p-p_k)\,\delta(x-x_k,\ y-y_k)\,\delta(\tau-\tau_k),$$
$$\tau_k = \left\lfloor \frac{t_k - t_a}{t_b - t_a} T \right\rfloor \tag{1}$$

State Space Models A classical linear time invariant state-space model treats each channel independently. For the input signal, it maps each channel of input $x(t)$ to the corresponding scalar channel output $y(t)$ through an N-dimensional latent state $h(t) \in \mathbb{C}^N$, where $\mathbf{A} \in \mathbb{C}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{C}^{1\times N}$, and $\mathbf{D} \in \mathbb{C}^{1\times 1}$ govern the system dynamics.

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t), \qquad y(t) = \mathbf{C}h(t) + \mathbf{D}x(t) \tag{2}$$

To operate on sampled sequences, the continuous parameters are converted to discrete ones using a time step Δ and a discretization rule such as zero-order hold:

$$\bar{A} = \exp(\Delta A), \quad \bar{B} = (\Delta A)^{-1}\big(\exp(\Delta A) - I\big)\,\Delta B. \tag{3}$$

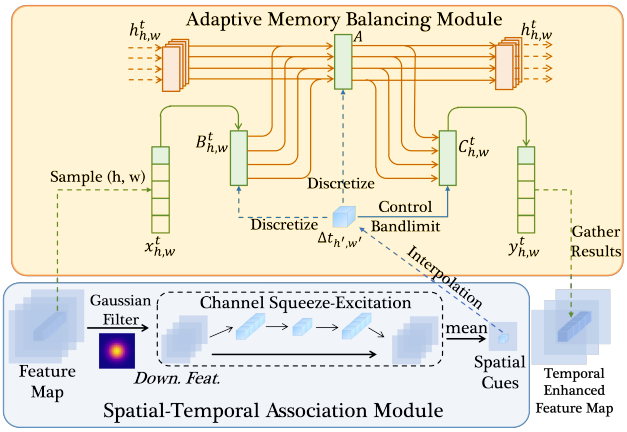

Figure 3: Internal view of the Spatial-Aware Temporal Module. The Spatial-Temporal Association Module extracts arbitrary-scale spatial-association cues with minimal computation and parameter overhead, then supplies Δt to steer the Adaptive Memory Balancing Module's state update.

The timescale Δ controls the temporal resolution of the system and plays a role analogous to gating in recurrent networks (Gu et al. 2020b; Tallec and Ollivier 2018).

Spatial-Temporal Association Module

We first build a Gaussian pyramid (Burt 1984) by iteratively down-sampling the feature map with a strided Gaussian kernel, suppressing high-frequency noise and aliasing while retaining spatial information as much as possible until the desired fusion scale k is reached. Then, a lightweight squeeze-and-excitation block (Hu, Shen, and Sun 2018) is adopted to model global channel interactions. A channel-wise mean followed by a sigmoid activation yields a saliency map of size $(1,\, H/k,\, W/k)$; this map quantifies information change within each $k \times k$ window and acts as a spatial-association cue for the temporal module.

Unlike traditional ConvLSTM or ConvGRU, whose parameter counts grow quadratically with receptive-field size, our module keeps its parameter budget independent of spatial extent. This design supports the efficient aggregation of spatial information at arbitrary, even global scales and passes these stable wide-area signals to the temporal module, encouraging robust memory updates.

Adaptive Memory Balancing Module

The Adaptive Memory Balancing Module builds on Mamba's efficient selective scan implementation (Gu and Dao 2023), inheriting its high-throughput training, low-latency inference, and long-range dependency modeling advantages. We further extend this mechanism by redesigning how the time step Δt is estimated and exploited.

In an SSM, with commonly used diagonal A matrix (Gu et al. 2022) and its n-th entry a_n, the hidden state of an SSM decomposes into independent components $h_n(t)$, and each basis function simplifies to $K_n(t) = e^{ta_n} B_n$. The imaginary part $\Im(a_n)$ controls its frequency (Nguyen et al. 2022), as shown in Eq. 4. If the sampling rate r is too low, high-frequency elements f_n exceed the Nyquist limit, causing aliasing and unpredictable state distortion. S4ND (Nguyen et al. 2022) and S5ViT (Zubic, Gehrig, and Scaramuzza 2024) mitigate this by scaling Δt with a global rate r: $r = 1$ during training, and $r = f_{\text{eval}}/f_{\text{train}}$ at the inference stage. f_{eval} and f_{train} represent time-domain training frequency and evaluation frequency. Then, they zero out coefficients C_n whose frequencies surpass $\alpha/2$ as depicted in Eq. 4. The hyper-parameter α acts as a band-limiting frequency threshold, with $\alpha = 1.0$ representing the Nyquist limit. Both methods suggest a lower empirical threshold, such as $\alpha = 0.5$.

$$f_n = \frac{\Delta t}{r} \cdot \frac{|\Im(a_n)|}{2\pi}, \quad C_n = \begin{cases} C_n, & \text{if } f_n \le \dfrac{\alpha}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Event-camera streams, however, exhibit drastic motion-dependent variability; relying on the time-domain frequency ratio r leaves spatial dynamics unmodeled. We therefore remove the parameter r and instead sample Δt from the motion-induced spatial indicator produced by the Spatial-Temporal Association Module. This Δt simultaneously guides discretization and bandlimiting: drastic changes in spatial semantics enlarge Δt, strengthen low-pass filtering, and bias updates toward the current input, whereas nearly-static scenes shrink Δt and preserve more historical information. It functions similarly to an RNN forget gate, yet with far fewer additional parameters and flexible spatial receptive fields. In summary, we leverage joint spatio-temporal cues to suppress aliasing and maintain robust, interpretable state evolution across both rapid-motion and static scenarios.

Event-oriented Data Augmentation

Real event streams contain far more random spikes than their simulated counterparts because of sensor noise. Beyond merely dropping events as in EventDrop (Gu et al. 2021), we also inject noise events at randomly sampled rates, thereby emulating hardware imperfections.

Simulated datasets also lack the motion diversity of real scenes. To unlock the full capacity of the Spatial-Temporal Association Module and Adaptive Memory Balancing Module, we introduce a suite of event-oriented spatio-temporal augmentations that synthesize richer motion patterns and sequence variations. These augmentations expose the model to a broader spatio-temporal distribution during training, markedly improving overall robustness and, in particular, the generalization ability of the temporal module.

Random Motion Pattern Generation We introduce a random non-uniform processing strategy on the initial event frame sequences to generate more diverse event frame sequences that represent varied motion patterns. Specifically, given an event frame sequence of length L, where each sequence element $E_e^{(i)}$ is a frame representation generated by Eq. 1 over consecutive, equally spaced intervals $[t^{(i)}, t^{(i+1)})$. We define a maximum acceleration factor R and randomly sample an acceleration ratio $r \in (1, R)$, yielding a new accelerated sequence length $M = \lfloor \frac{L}{r} \rfloor$. We

then select $M-1$ strictly increasing indices from the set $1,\ldots,L-1$.

$$0 = s_0 < s_1 < \ldots < s_{M-1} < s_M = L. \quad (5)$$

The sampled indices are used to merge the original tensor into $\tilde{E}_{\mathcal{E}}^{(m)}$. Unlike the original uniform intervals $\Delta t = t^{n+1} - t^n$, each merged tensor represents events occurring within non-uniform time intervals $\Delta t_m = t^{s_{m+1}} - t^{s_m}$. This effectively simulates frames generated by complex random non-uniform acceleration of the original frame sequence.

Besides merging for acceleration, we further simulate low-speed motion using magnitude-aware stochastic dropping. Specifically, we randomly sample a drop ratio β within the range $(0,1)$. Sampling weights w_{x_k,y_k} are computed based on the frequency of events at pixel locations (x_k,y_k) over the time interval. Importance-based sampling selects an event subset D, discarding it to form $\tilde{E}_{\mathcal{E}\setminus\mathcal{D}}^{(n)}$. This lightweight augmentation both imitates low-velocity motion and suppresses overly dense regions, guiding the model to focus on sparser spatial cues.

Priority-based Masked Sequence Modeling In real-world congested or static scenes, an event camera may record almost no events for extended periods, whereas in the simulation dataset, almost every training frame contains abundant events for spatial feature extraction, leaving the temporal module under-utilized. To compel the network to exploit temporal cues under sparse event streams, we propose Priority-based Spatial Masking, inspired by masked modeling techniques such as MAE (Feichtenhofer et al. 2022) and BEiT (Bao et al. 2021).

The image plane is partitioned into $a \times b$ non-overlapping patches of size $n \times n$:

$$\mathcal{P}_{a,b} = \{(x,y) \mid \lfloor x/n \rfloor = a,\ \lfloor y/n \rfloor = b\}. \quad (6)$$

For a sequence of length L, we accumulate the events in each patch as depicted in Eq. 7 and convert the counts into importance weights $q_{a,b}$.

$$C_{a,b} = \sum_{n=0}^{L-1} \sum_{e_k \in E_{\mathcal{E}}^{(n)}} \delta_{\mathcal{P}_{a,b}}(x_k,y_k) \quad (7)$$

$$q_{a,b} = \frac{\exp(C_{a,b})}{\sum_{a',b'} \exp(C_{a',b'})} \quad (8)$$

A proportion $\alpha \in (0,1)$ of spatial patches is selected by importance sampling according to $q_{a,b}$ to form the spatial mask $\mathcal{M}$. Independently, a proportion $\beta \in (0,1)$ of frames is drawn uniformly from the sequence, and $\mathcal{M}$ is applied to the selected frames. The network is therefore forced to predict the depths of the masked regions from the remaining spatio-temporal context.

Temporal random sampling prevents the recurrent state from overfitting to specific frame indices, while spatial masking selects sequence-level salient patches, ensuring the remaining frames still provide complementary cues for reconstructing depth in masked regions. Combined, these strategies compel the model to exploit cross-frame context, enhancing the temporal module's robustness.

Experiment

Experiment Setup

We examine the zero-shot relative-depth performance of models trained only on synthetic data when tested on out-of-domain real-world event datasets.

Datasets & Evaluation Metrics Training and validation are conducted on EventScape (Gehrig et al. 2021a), a CARLA-based simulation suite with 743 driving sequences and 171,000 densely annotated frames, which is markedly larger than the existing real-world counterpart and provides pixel-accurate depth. Zero-shot testing is performed on the widely used real-world DSEC (Gehrig et al. 2021b) and MVSEC (Zhu et al. 2018); train/val/test splits follow the protocol of (Hidalgo-Carrió, Gehrig, and Scaramuzza 2020; Gehrig et al. 2021a) and are reported in the appendix.

Prior event-based methods define dataset-specific logarithmic depth scale parameters from dataset statistics (Hidalgo-Carrió, Gehrig, and Scaramuzza 2020; Gehrig et al. 2021a; Liu et al. 2024b), limiting cross-dataset transfer. Following MiDaS (Ranftl et al. 2020), we perform per-frame scale-and-shift alignment before computing the evaluation metrics: absolute relative error ($Abs.Rel$), logarithmic mean squared error ($RMSE_{Log}$), scale invariant logarithmic error ($SILog$), and accuracy $\delta < 1.25^n$ (n=1, 2, 3). Detailed metric definitions are provided in the appendix.

Training Loss & Network Architecture Our training loss mirrors MiDaS (Ranftl et al. 2020): a scale- and shift-invariant term plus a multi-scale, scale-invariant gradient loss. The framework is agnostic to the spatial backbone; for computing efficiency and global receptive field we adopt MobileMamba (He et al. 2025) as the spatial extractor. The network starts with a convolutional patch-embedding layer (Zhang et al. 2023; He et al. 2025) that produces spatial tokens. These tokens, together with the recurrent hidden states, are processed by a U-Net with skip connections, yielding a full-resolution depth map.

Comparison with the State-of-the-Art

Table 1 reports the comparative results. We select two representative open-source event-based monocular depth estimation methods, MDDE (Hidalgo-Carrió, Gehrig, and Scaramuzza 2020) and EReFormer (Liu et al. 2024b), as baselines. In addition, we adapt encoders from three state-of-the-art event-based object detectors, RVT (Gehrig and Scaramuzza 2023), SSM-E (Zubic, Gehrig, and Scaramuzza 2024), and SMamba (Yang et al. 2025), to depth estimation, retraining them with our loss function to ensure a fair comparison. Evaluations include in-domain testing on EventScape and zero-shot transfer to two common real-world datasets, MVSEC and DSEC. On EventScape, our MA-Mamba achieves state-of-the-art performance with competitive parameter counts and GFLOPs.

On zero-shot evaluation with real-world datasets, SMamba (Yang et al. 2025) is an exception: its Spatial-Channel Interaction Mixing module is sensitive to resolution variations, leading to inconsistent cross-dataset behavior.

Dataset	Method	Abs. Rel ↓	RMSE_{Log} ↓	SILog ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	FLOPs ↓	Params ↓	FPS ↑
EventScape (in-domain)	MDDE	0.18	**0.31**	**0.10**	0.78	**0.90**	0.94	122.18	10.7M	113.32
	EReFormer	0.21	0.35	0.16	0.75	0.88	0.91	132.64	89.5M	20.01
	RVT*	0.21	0.40	0.15	0.68	0.86	0.93	14.70	13.7M	105.6
	SSM-E*	0.18	0.36	0.13	0.74	0.87	0.93	**3.71**	**3.4M**	141.08
	SMamba*	0.17	**0.31**	**0.10**	**0.79**	**0.90**	**0.95**	13.16	11.2M	18.35
	MA-Mamba	**0.16**	0.32	**0.10**	0.77	0.89	**0.95**	11.35	4.0M	**220.75**
MVSEC (zero-shot)	MDDE	0.41	0.54	0.27	0.43	0.67	0.81	114.54	10.7M	120.49
	EReFormer	0.64	0.67	0.42	0.30	0.54	0.70	116.76	89.5M	20.34
	RVT*	0.37	0.48	0.21	0.47	0.70	0.84	13.56	13.7M	106.93
	SSM-E*	0.48	0.51	0.24	0.44	0.65	0.79	**3.35**	**3.4M**	142.0
	SSM-E+	0.44	0.50	0.22	0.46	0.63	0.84	**3.35**	**3.4M**	142.0
	SMamba*	0.49	0.53	0.27	0.40	0.66	0.82	36.32	11.2M	9.99
	MA-Mamba	**0.31**	**0.42**	**0.17**	**0.54**	**0.77**	**0.89**	10.65	4.0M	**234.7**
DSEC (zero-shot)	MDDE	0.35	0.44	0.19	0.42	0.71	0.88	305.45	10.7M	49.07
	EReFormer	0.34	0.47	0.23	0.43	0.74	0.90	387.63	89.5M	11.60
	RVT*	0.29	0.37	0.17	0.51	0.82	0.91	37.15	13.7M	54.41
	SSM-E*	0.32	0.31	0.22	0.46	0.77	0.88	**9.40**	**3.4M**	72.29
	SSM-E+	0.40	0.46	0.26	0.33	0.62	0.85	**9.40**	**3.4M**	72.29
	SMamba*	0.44	0.54	0.27	0.28	0.54	0.76	39.48	11.2M	6.64
	MA-Mamba	**0.20**	**0.27**	**0.07**	**0.67**	**0.90**	**0.98**	28.39	4.0M	**112.38**

Table 1: Comparison of SOTA methods evaluated in-domain on EventScape and zero-shot on MVSEC and DSEC. An asterisk (*) denotes models retrained for depth estimation; a plus (+) indicates *SSM-E* with its rate adjusted to each dataset's inference frequency. Metrics marked with ↓ are better when lower, and ↑ when higher; the best result is in **bold**, second best is underlined. FPS is measured on a single NVIDIA A100 GPU (batch size 1). Our proposed MA-Mamba consistently achieves top robustness on real-world datasets with high efficiency.

All other methods maintain stable relative rankings, alleviating dataset bias concerns. SSM-E (Zubic, Gehrig, and Scaramuzza 2024) offers a tunable rate parameter to account for different time-domain inference frequencies of different datasets, but this manual adjustment can degrade performance in cross-dataset settings, particularly on DSEC. In contrast, our model achieves the best zero-shot results on both MVSEC and DSEC without any explicit parameter tuning, substantially outperforming the second-best method and demonstrating its superior adaptability and robustness.

Ablation Studies

Because of space constraints, we report only three metrics: $Abs.Rel$, $RMSE_{Log}$, and $SILog$ for the zero-shot evaluation on MVSEC. Similar performance trends are observed on DSEC. More results are available in the Appendix.

Effectiveness of the Proposed Modules In Table 2, we evaluate each component's contribution to cross-dataset robustness: Spatial-Temporal Association (STA), Adaptive Memory Balancing (AMB), and event-oriented data augmentations (EDA). Each module yields improvements on its own. However, AMB alone yields only a modest improvement; we infer that without STA it receives noisy and insufficient spatial cues. EDA alone produces a notable boost; when combined with STA and AMB, it exposes the model to richer spatio-temporal patterns, facilitating further robustness improvement.

STA	AMB	EDA	Abs. Rel ↓	RMSE_{Log} ↓	SILog ↓
			0.50	0.49	0.23
✓			0.45	0.47	0.21
	✓		0.47	0.49	0.22
		✓	0.44	0.46	0.21
✓	✓		0.40	0.45	0.19
✓	✓	✓	0.31	0.42	0.17

Table 2: Ablation results under MVSEC zero-shot testing. "✓" indicates the presence of the component.

Noise	Masking	Motion	Abs. Rel	RMSE_{Log}	SILog
			0.40	0.45	0.19
✓			0.38	0.45	0.19
✓	✓		0.35	0.44	0.18
✓	✓	✓	0.31	0.42	0.17

Table 3: Ablation of data augmentation strategies under MVSEC zero-shot testing. "Noise" represents Random Noise Injection, "Masking" represents Priority-based Masked Sequence Modeling, and "Motion" represents Random Motion Pattern Generation.

Effect of Individual Augmentations We conduct ablations on Random Noise Injection, Random Motion Pattern Generation, and Priority-based Masked Sequence Modeling to evaluate their individual contributions. As illustrated in Tab. 3, the latter two augmentations offer notable perfor-

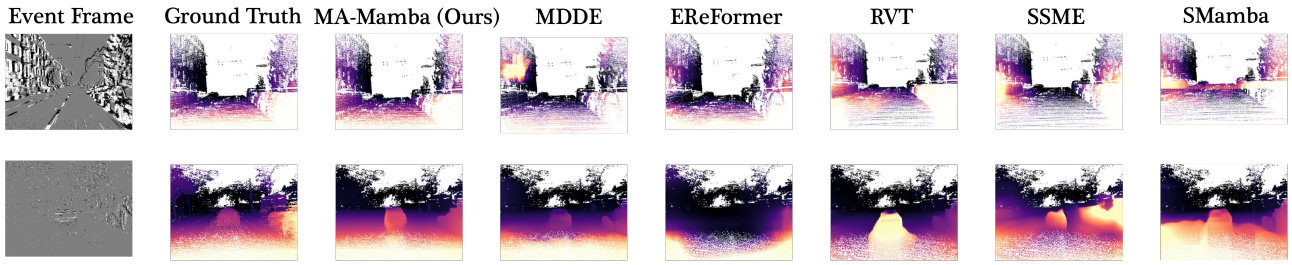

Figure 4: Zero-shot qualitative comparisons on real-world datasets. The top row shows results on DSEC, and the bottom row shows results on MVSEC.

mance improvements, whereas random noise injection contributes little. This indicates that the spatial feature extractor might already be robust to noise, and the primary bottleneck lies in the temporal module, precisely the main focus of our paper.

Receptive Field	Abs. Rel	RMSE_{Log}	SILog
8×8	0.37	0.44	0.20
16×16	0.34	0.43	0.18
32×32	0.31	0.42	0.17
Global	0.32	0.42	0.16

Table 4: Effect of STA receptive-field size in MVSEC zero-shot testing. "Global" denotes a single spatial indicator shared across all locations.

Granularity of the STA Module Tab. 4 examines the effect of STA's spatial granularity. An overly small receptive field fails to provide informative spatial indicators, negatively affecting performance. As granularity increases to a moderate level, performance differences narrow. Beyond that point, approaching too large and even global granularity, the model becomes less sensitive to meaningful local variations, leading to another slight decline.

Spatial Feature Extractor	Abs. Rel	RMSE_{Log}	SILog	Params	GFLOPs
EfficientNet-V3	0.34	0.44	0.20	3.3M	5.32
EfficientViT-M1	0.33	0.42	0.18	4.0M	4.22
MobileMamba-T2	0.31	0.42	0.17	4.0M	10.65

Table 5: Impact of substituting the spatial feature extractor on MVSEC zero-shot performance.

Choice of Spatial Feature Extractor Our framework is compatible with various spatial backbones. As shown in Tab. 5, we comprehensively evaluate compact mobile convolutional (Howard et al. 2019), attention-based (Liu et al. 2023), and Vision-Mamba variants (He et al. 2025). After considering computational efficiency and performance, we select MobileMamba (He et al. 2025) as our spatial feature extractor due to its favorable trade-off.

Data Scaling Experiment To examine how additional synthetic data affect performance, we enlarged the training set by appending the EventScape validation split and the DENSE dataset (Hidalgo-Carrió, Gehrig, and Scaramuzza 2020), increasing the total volume to roughly $1.5\times$ the original EventScape training set. We benchmarked our model against RVT (Gehrig and Scaramuzza 2023), the second-best performer on MVSEC. As depicted in Tab. 6, with the same data-scaling factor, our method achieves a larger gain from scaling data in robustness, suggesting considerable headroom as more synthetic data become available.

Training Set	Method	Abs. Rel	RMSE_{Log}	SILog
Standard	RVT	0.37	0.48	0.21
	MA-Mamba	0.31	0.42	0.17
Enlarged	RVT	0.35	0.46	0.20
	MA-Mamba	0.27	0.39	0.15

Table 6: Experiments on more synthetic training data.

Qualitative Results

Fig. 4 shows visualizations on challenging MVSEC and DSEC samples. On DSEC, only our model recovers accurate relative depths for cars, trees, and buildings in regions where LiDAR ground truth is incomplete, while others produce spurious artifacts that we infer arise from untimely memory forgetting. In MVSEC, under weak event signals around the central car, other methods either miss it or exhibit boundary bleeding; our approach, with more robust memory, preserves its position and sharp contours.

Conclusion

We introduced MA-Mamba, a lightweight yet robust architecture that adapts to unseen motion patterns and narrows the sim-to-real gap in event-based monocular depth estimation through motion-aware architecture design and event-oriented augmentation. The model delivers impressive zero-shot quantitative and qualitative results on MVSEC and DSEC, and data-scaling experiments suggest further gains with larger synthetic corpora. Our framework paves the way for robust, low-power 3D perception under all lighting conditions in autonomous driving and robotics.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China Grant 62373315, U24A20252,

U23A20339, National Key Research and Development Program of China Grant 2024YFB4707603, Guangzhou Municipal Project 2023A03J0011 and 2024D03J0008, Guangdong Provincial Project 2023ZDZX1037 and 2023ZT10X009, and the Nansha Key Science and Technology Project 2023ZD006.

References

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Burt, P. J. 1984. The pyramid as a structure for efficient computation. In *Multiresolution image processing and analysis*, 6–35. Springer.

Cordone, L.; Miramond, B.; and Thierion, P. 2022. Object detection with spiking neural networks on automotive event data. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Dong, X.; Garratt, M. A.; Anavatti, S. G.; and Abbass, H. A. 2022. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(10): 16940–16961.

Duan, P.; Wang, Z. W.; Zhou, X.; Ma, Y.; and Shi, B. 2021. EventZoom: Learning to denoise and super resolve neuromorphic events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12824–12833.

Feichtenhofer, C.; Li, Y.; He, K.; et al. 2022. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35: 35946–35958.

Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1): 154–180.

Gehrig, D.; Rüegg, M.; Gehrig, M.; Hidalgo-Carrió, J.; and Scaramuzza, D. 2021a. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2): 2822–2829.

Gehrig, M.; Aarents, W.; Gehrig, D.; and Scaramuzza, D. 2021b. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3): 4947–4954.

Gehrig, M.; and Scaramuzza, D. 2023. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13884–13893.

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

Gu, A.; Dao, T.; Ermon, S.; Rudra, A.; and Ré, C. 2020a. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33: 1474–1487.

Gu, A.; Goel, K.; Gupta, A.; and Ré, C. 2022. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35: 35971–35983.

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

Gu, A.; Gulcehre, C.; Paine, T.; Hoffman, M.; and Pascanu, R. 2020b. Improving the gating mechanism of recurrent neural networks. In *International conference on machine learning*, 3800–3809. PMLR.

Gu, F.; Sng, W.; Hu, X.; and Yu, F. 2021. Eventdrop: Data augmentation for event-based learning. *arXiv preprint arXiv:2106.05836*.

He, H.; Zhang, J.; Cai, Y.; Chen, H.; Hu, X.; Gan, Z.; Wang, Y.; Wang, C.; Wu, Y.; and Xie, L. 2025. Mobilemamba: Lightweight multi-receptive visual mamba network. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4497–4507.

Hidalgo-Carrió, J.; Gehrig, D.; and Scaramuzza, D. 2020. Learning monocular dense depth from events. In *2020 International Conference on 3D Vision (3DV)*, 534–542. IEEE.

Howard, A.; Sandler, M.; Chu, G.; Chen, L.-C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1314–1324.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hu, M.; Yin, W.; Zhang, C.; Cai, Z.; Long, X.; Chen, H.; Wang, K.; Yu, G.; Shen, C.; and Shen, S. 2024. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kugele, A.; Pfeil, T.; Pfeiffer, M.; and Chicca, E. 2023. How many events make an object? improving single-frame object detection on the 1 mpx dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3913–3922.

Liu, F.; Huang, T.; Zhang, Q.; Yao, H.; Zhang, C.; Wan, F.; Ye, Q.; and Zhou, Y. 2024a. Ray denoising: Depth-aware hard negative sampling for multi-view 3d object detection. In *European Conference on Computer Vision*, 200–217. Springer.

Liu, X.; Li, J.; Fan, X.; and Tian, Y. 2022. Event-based Monocular Dense Depth Estimation with Recurrent Transformers. *arXiv preprint arXiv:2212.02791*.

Liu, X.; Li, J.; Shi, J.; Fan, X.; Tian, Y.; and Zhao, D. 2024b. Event-based monocular depth estimation with recurrent transformers. *IEEE Transactions on Circuits and Systems for Video Technology*.

Liu, X.; Peng, H.; Zheng, N.; Yang, Y.; Hu, H.; and Yuan, Y. 2023. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14420–14430.

Nguyen, E.; Goel, K.; Gu, A.; Downs, G.; Shah, P.; Dao, T.; Baccus, S.; and Ré, C. 2022. S4nd: Modeling images and videos as multidimensional signals with state spaces. *Advances in neural information processing systems*, 35: 2846–2861.

Pan, T.; Cao, Z.; and Wang, L. 2024. Srfnet: Monocular depth estimation with fine-grained structure via spatial reliability-oriented fusion of frames and events. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 10695–10702. IEEE.

Peng, L.; Chen, Z.; Fu, Z.; Liang, P.; and Cheng, E. 2023. Bevsegformer: Bird's eye view semantic segmentation from arbitrary camera rigs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5935–5943.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3): 1623–1637.

Rebecq, H.; Ranftl, R.; Koltun, V.; and Scaramuzza, D. 2019. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6): 1964–1980.

Ren, H.; Zhou, Y.; Zhu, J.; Lin, X.; Fu, H.; Huang, Y.; Fang, Y.; Ma, F.; Yu, H.; and Cheng, B. 2025. Rethinking efficient and effective point-based networks for event camera classification and regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ren, H.; Zhu, J.; Zhou, Y.; Fu, H.; Huang, Y.; and Cheng, B. 2024. A simple and effective point-based network for event camera 6-dofs pose relocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18112–18121.

Schaefer, S.; Gehrig, D.; and Scaramuzza, D. 2022. Aegnn: Asynchronous event-based graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12371–12381.

Shen, G.; Zhao, D.; and Zeng, Y. 2023. Eventmix: An efficient data augmentation strategy for event-based learning. *Information Sciences*, 644: 119170.

Sun, M.; Zhang, D.; Ge, Z.; Wang, J.; Li, J.; Fang, Z.; and Xu, R. 2024. Eventrpg: Event data augmentation with relevance propagation guidance. *arXiv preprint arXiv:2403.09274*.

Tallec, C.; and Ollivier, Y. 2018. Can recurrent neural networks warp time? *arXiv preprint arXiv:1804.11188*.

Tian, Y.; Chen, H.; Deng, Y.; Shen, F.; Liu, K.; You, W.; and Zhang, Z. 2024. EventAug: Multifaceted Spatio-Temporal Data Augmentation Methods for Event-based Learning. *arXiv preprint arXiv:2409.11813*.

Wang, Y.; Chao, W.-L.; Garg, D.; Hariharan, B.; Campbell, M.; and Weinberger, K. Q. 2019. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8445–8453.

Yang, C.; Chen, Y.; Tian, H.; Tao, C.; Zhu, X.; Zhang, Z.; Huang, G.; Li, H.; Qiao, Y.; Lu, L.; et al. 2023. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17830–17839.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10371–10381.

Yang, N.; Wang, Y.; Liu, Z.; Li, M.; An, Y.; and Zhao, X. 2025. SMamba: Sparse Mamba for Event-based Object Detection. *arXiv preprint arXiv:2501.11971*.

Zhang, C.; Han, D.; Qiao, Y.; Kim, J. U.; Bae, S.-H.; Lee, S.; and Hong, C. S. 2023. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*.

Zhang, J.; Tang, L.; Yu, Z.; Lu, J.; and Huang, T. 2022. Spike transformer: Monocular depth estimation for spiking camera. In *European Conference on Computer Vision*, 34–52. Springer.

Zhu, A. Z.; Thakur, D.; Özaslan, T.; Pfrommer, B.; Kumar, V.; and Daniilidis, K. 2018. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robotics and Automation Letters*, 3(3): 2032–2039.

Zubic, N.; Gehrig, M.; and Scaramuzza, D. 2024. State space models for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5819–5828.