E-MaT: Event-Oriented Mamba for Egocentric Point Tracking

Han Han[1], Wei Zhai[1]*, Baocai Yin[2], Yang Cao[1], Bin Li[1], Zheng-jun Zha[1]

[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China
[2]iFlytek Research, iFlytek Co., Ltd., Hefei, Anhui, China

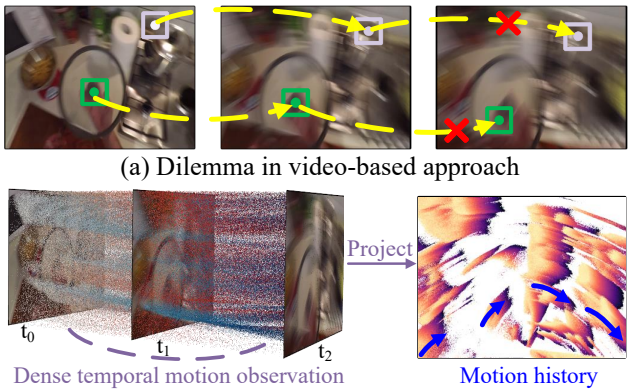
hanh@mail.ustc.edu.cn, wzhai056, forrest, binli, zhazj@ustc.edu.cn, bcyin@iflytek.com

Abstract

Egocentric point tracking aims to localize points on object surfaces from a first-person perspective and serves as a critical step toward embodied intelligence. Recent methods rely on video input, tracking query points through feature matching across consecutive frames. However, these methods struggle in highly dynamic settings—a common challenge in first-person perspectives, where the head-mounted camera undergoes frequent and abrupt rotations, resulting in high angular velocities, motion blur, and large inter-frame displacements. In contrast, event cameras capture motion at microsecond temporal resolution, naturally avoiding blur and delivering low-latency, high-fidelity cues crucial for egocentric point tracking. Moreover, rapid egocentric motion disrupts local smoothness, breaking the assumption that spatially adjacent regions share similar motion. Event dynamics expose global motion trends, guiding coherent modeling and consistent feature flow. Therefore, this paper proposes a mamba-based tracking framework that constructs feature modeling paths aligned with the dominant motion trend extracted from events, and modulates feature propagation along these paths based on local motion intensity, enhancing stability by suppressing unreliable signals and emphasizing consistent cues. Additionally, a motion-adaptive suppression module enhances temporal robustness by adaptively suppressing correlation features based on motion intensity variations, mitigating the effects of intensity fluctuations and partial observability. To facilitate research in this domain, a multimodal dataset named DVS-EgoPoints with both events and videos for egocentric point tracking is collected. Experiments on the DVS-EgoPoints dataset and a simulation benchmark demonstrate superior performance over state-of-the-art methods, especially under challenging motion and occlusion conditions.

Introduction

Egocentric point tracking aims to track points in a scene from a first-person perspective. It is a crucial component of embodied AI (Plizzari et al. 2024) and has applications in robotics (Behrens et al. 2025), as well as AR and VR (Zhao et al. 2025). Recent methods (Karaev et al. 2024; Kim et al. 2025) rely on video input, extracting per-frame features via an encoder and matching the query region over time. While effective in third-person tracking, these approaches struggle in first-person settings. Third-person views typically exhibit smooth, constrained camera motion, with objects largely remaining in view. In contrast, egocentric data from head-mounted cameras involves abrupt head movements, causing motion blur and large inter-frame displacements that degrade appearance-based matching accuracy, see Figure 1(a).



(a) Dilemma in video-based approach

(b) Advantage of event-based approach

Figure 1: (a) Video-based methods primarily track query points via appearance matching, which fails under fast camera motion due to motion blur. (b) Event cameras provide dense temporal observations without motion blur. In addition, event projections reflect both spatial structure and motion evolution, with darker colors marking earlier timestamps and color transitions encoding motion history, enabling motion-aware tracking.

To address these challenges, this paper leverages event cameras (Lichtsteiner, Posch, and Delbruck 2008; Taverni et al. 2018), where each pixel independently responds to brightness changes at microsecond-level time resolution, avoiding motion blur. They reveal more dynamic scene changes rather than appearance information, making them an effective alternative to video in motion-focused tasks. Furthermore, the event order implicitly carries motion cues, providing guidance for motion-aware feature modeling, as shown in Figure 1(b).

However, event triggering frequency at each pixel en-

*Corresponding author.

codes not only spatial structure but also temporal motion intensity. This spatiotemporal coupling causes existing methods, which primarily focus on spatial encoding, to lose information. Moreover, these methods often assume local spatial-motion consistency: CNNs aggregate over local neighborhoods, Transformers split inputs into adjacent patches, and Mamba follows a fixed scan path. Under rapid egocentric motion, such assumptions fail, leading to misaligned feature aggregation and degraded temporal modeling.

Therefore, this paper presents an event-based egocentric point tracking framework designed to capture spatiotemporal dependencies in event data. Specifically, an event-guided mamba encoder is proposed to construct feature modeling paths aligned with the dominant motion trend extracted from events, enabling spatially consistent and temporally aware representation. Feature propagation along these paths is further modulated by local motion intensity to suppress unreliable signals and emphasize coherent cues. Additionally, a Motion-Adaptive Suppression (MAS) module is introduced to enhance robustness under rapid motion and partial occlusion by dynamically suppressing correlation features in regions with high motion intensity variations. To validate the effectiveness of this method, the paper collects an egocentric point tracking dataset named DVS-EgoPoints, which includes both event and video modalities. The proposed method shows superior performance on both the DVS-EgoPoints and a simulation dataset. The main contributions of this paper can be summarized as follows:

- An event-based egocentric tracking framework named E-MaT is proposed, capable of tracking fast-moving points across occlusions in first-person scenarios.
- An event-guided Mamba encoder that adapts the feature modeling path to the dominant motion trend and modulates feature propagation by local motion intensity, enabling coherent spatiotemporal representation.
- A motion-adaptive suppression module complements the encoder by attenuating noisy correlations under rapid motion and occlusion, stabilizing feature matching.
- For comprehensive evaluation, a real-world egocentric point tracking dataset with both events and videos is collected. Experiments on synthetic and real-world datasets reveal the competitive performance of E-MaT.

Related Work

Point Tracking

Point tracking is a fundamental task in computer vision, aiming to determine the subsequent positions of a given query point on a physical surface over time. Most existing methods rely on video input, extracting appearance features around the query point in the initial frame and searching for the best-matching region in subsequent frames (Doersch et al. 2022; Karaev et al. 2023; Cho et al. 2024; Qu et al. 2024; Karaev et al. 2024; Li et al. 2025). For example, the PIPs series (Harley, Fang, and Fragkiadaki 2022; Zheng et al. 2023) employs ResNet (He et al. 2016) to extract frame-wise features and performs temporal matching within the query's neighborhood to track point positions. These approaches achieve impressive performance in third-person videos where motion is smooth and appearance remains relatively stable.

In contrast, egocentric settings present unique challenges due to rapid head movements, motion blur, and frequent viewpoint changes, all of which undermine the stability of appearance-based cues. To investigate this, (Darkhalil et al. 2024) introduces two egocentric point tracking datasets: K-EPIC-Point for training and EgoPoints for evaluation, and benchmarks existing trackers in this context. Their performance noticeably diverges from third-person scenarios, revealing limitations in handling egocentric dynamics.

Event-based Vision

Event cameras (Lichtsteiner, Posch, and Delbruck 2008) are bio-inspired vision sensors that asynchronously capture per-pixel brightness changes with microsecond latency. Compared to conventional cameras, they offer higher dynamic range, negligible motion blur, and ultra-low power consumption. These advantages have led to growing interest in event-based vision tasks such as optical flow estimation (Wan et al. 2024; Shiba et al. 2024; Wan et al. 2025; Almatrafi et al. 2020; Liao et al. 2024), action recognition (Wan et al. 2022; Tan et al. 2022; Gao et al. 2023), and point tracking (Liu et al. 2024a; Hamann et al. 2025; Han et al. 2025). Existing event-based point tracking methods primarily focus on third-person scenarios, following pipelines similar to video-based approaches: spatial features are extracted using encoder backbones, while temporal dependencies are modeled subsequently in a separate stage. For example, FE-TAP (Liu et al. 2024a) combines spatial features from events and videos and employs temporal attention for point tracking. However, such methods assume local motion smoothness and spatial continuity, which fails under rapid egomotion, limiting their effectiveness in first-person scenarios.

State Space Model

State Space Models (SSMs), originally from control theory, have shown promise in modeling long-range dependencies via linear recurrent dynamics. Mamba (Gu and Dao 2023) enhances SSMs with input-dependent, time-varying parameters, achieving global context modeling with linear complexity. Unlike ViT's global self-attention (Dosovitskiy et al. 2021), Mamba processes image patches sequentially, enabling flexible scan directions. Variants like Vim (Zhu et al. 2024), VMamba (Liu et al. 2024b), and Mamba-ND (Li, Singh, and Grover 2025) explore different spatial scan strategies to improve efficiency and context integration.

Recent studies (Huang et al. 2024; Ge et al. 2025; Wang et al. 2024; Zubic, Gehrig, and Scaramuzza 2024) have increasingly explored mamba models for event data, yet few have examined how the spatiotemporal characteristics of events influence feature extraction. Unlike video, where temporal dependencies span across frames and intra-frame content is spatial, event data simultaneously captures both scene structure and precise temporal dynamics. This encoding presents a challenge for mamba models with fixed scan directions, which traditionally focus on spatial modeling and overlook the temporal cues inherent in event data.

Method

Event Processing

Event cameras asynchronously emit events when the change in logarithmic intensity exceeds a threshold C:

$$\log I(x, y, t) - \log I(x, y, t - \Delta t) = pC, \tag{1}$$

where Δt is the time between consecutive events, and $p \in \{-1, 1\}$ indicates polarity. Each event is a four-tuple (x, y, t, p), encoding the location, timestamp, and polarity.

To process events, this work adopts the time surface (Mueggler, Bartolozzi, and Scaramuzza 2017), where each pixel stores the timestamp of its most recent event within a temporal window. This window determines how far into the past events are considered when constructing the surface, enabling control over the temporal context. Unlike voxel grids (Zhu et al. 2019) or event frames (Liu and Delbruck 2018), the time surface retains fine-grained motion history (Gallego et al. 2022), making it well-suited for point tracking.

Framework

The problem can be defined as follows: given an event stream, which is mapped to a sequence of time surface representations $\mathcal{E} = \{\mathbf{E}_0, \mathbf{E}_1, ..., \mathbf{E}_T\} \in \mathbb{R}^{T\times 2\times H\times W}$, and an query point $x_{src} = x_0 \in \mathbb{R}^2$ at time step 0, the goal is to estimate its position $\mathbf{X} = \{x_0, x_1, ..., x_T\} \in \mathbb{R}^{T\times 2}$ and visibility $\mathbf{V} = \{v_0, ..., v_T\} \in \mathbb{R}^{T\times 1}$ across all time steps.

The proposed E-MaT framework consists of two stages: initialization and iterative refinement, with the overall network architecture illustrated in Figure 2. In the initialization stage, spatial-temporal features $\mathcal{F} = \{\mathbf{F}_0, \mathbf{F}_1, ..., \mathbf{F}_T\} \in \mathbb{R}^{T\times C\times H\times W}$ are extracted from time surfaces using an Event-Oriented Mamba encoder. Unlike prior works (Karaev et al. 2023; Zheng et al. 2023) that focus solely on spatial features at individual time steps, EOM leverages motion history encoded in the time surface to extract temporally aware spatial features, enhancing feature expressiveness.

During the iterative refinement phase, let $x_t^k \in \mathbb{R}^2$ and $v_t^k \in \mathbb{R}^1$ denote the estimated point and visibility at time t after the k-th iteration. For each timestep t, E-MaT extracts multi-scale features $f_t \in \mathbb{R}^{4C\times 7\times 7}$ by applying average pooling to $\mathbf{F}_t$ at three scales and sampling 7×7 local regions around x_t. Correlation volumes $c_t \in \mathbb{R}^{7\times 7\times 7\times 7}$ are then computed between f_t and the query-step features. The resulting correlations $\mathbf{C}^k = \{c_0^k, ..., c_T^k\}$, together with the time surface $\mathcal{E}$ and positions $\mathbf{X}^k$, are refined by the MAS module, which attenuates unreliable correlations based on local motion intensity variations. This suppresses noisy activations caused by non-uniform event rates or occlusions, encouraging temporal consistency. The refined correlations $\hat{\mathbf{C}}^k$, positions $\mathbf{X}^k$, and visibility $\mathbf{V}^k$ are passed into a Transformer-based decoder (Karaev et al. 2024) to predict updates $\Delta\mathbf{X}$ and $\Delta\mathbf{V}$. Estimates are updated via: $\mathbf{X}^{k+1} = \mathbf{X}^k + \Delta\mathbf{X}$ and $\mathbf{V}^{k+1} = \mathbf{V}^k + \Delta\mathbf{V}$.

Event-oriented Mamba

The frame-like synchronous time surface representation encodes motion history (Gallego et al. 2022), where temporal relationships exist not only across time steps but also within a single frame. This contrasts with video, where spatial structures are captured within frames and temporal dependencies occur between frames. Therefore, video-based methods that first extract spatial features and then model temporal dependencies may cause loss of intra-frame temporal information in time surface, which is critical for point tracking. To tackle this, this paper proposes Event-Oriented Mamba (EOM) to exploit temporal motion information embedded within the time surface.

As shown in Figure 3, the input is first divided into non-overlapping patches via a patch embedding layer, followed by L stacked EOM layers to extract features. For clarity, multi-stage details are omitted in the figure and provided in the supplementary. The final output is a feature map $\mathbf{F}_t \in \mathbb{R}^{C\times H\times W}$. This operation is independently applied at each time step, allowing the model to adaptively capture the dominant motion direction over time.

Each EOM layer includes linear projections, normalization, depthwise convolutions (DWConv), and the key motion-guided scan module. Previous methods rely on fixed scan directions to model dependencies, which is suboptimal for event data, as it disrupts the temporal relationship in the time surface during sequence modeling. Motion-guided scan module determines the dominant motion direction by convolving the time surface with 8 directional Sobel operators, producing a gradient map $\mathbf{G}_t \in \mathbb{R}^{8\times H\times W}$. It then performs Global Average Pooling (GAP) to obtain a direction-wise motion response vector $w \in \mathbb{R}^8$, which is normalized via softmax:

$$w_k = \frac{\exp\left(\frac{1}{HW}\sum \mathbf{G}_t(k, x, y)\right)}{\sum_{j=1}^{8}\exp\left(\frac{1}{HW}\sum \mathbf{G}_t(j, x, y)\right)} \tag{2}$$

The scan direction with the highest response is selected as the dominant motion θ, and feature propagation is performed exclusively along this path to preserve temporal coherence.

In addition to directional selection, the local motion magnitude, computed from gradient responses within each patch, is used to reweight features. This modulation suppresses unreliable signals from noisy or static regions while emphasizing motion-consistent patches. Compared to fixed scanning, this adaptive mechanism reduces redundancy and enhances feature sensitivity to dynamic structures.

Motion-adaptive Suppression

Event cameras respond to motion only, with uniform motion resulting in a steady event-triggering frequency, while non-uniform motion, such as acceleration or deceleration, causes fluctuations in event-triggering frequency, which increase as the motion becomes more intense. These fluctuations can cause temporal inconsistencies and disrupt the matching process, leading to unstable feature correlations across time steps. To address this, the Motion-Aware Suppression (MAS) module suppresses correlation features from time steps with significant motion changes, leveraging motion variations to force the model to infer positions from more stable local time steps. Additionally, MAS alleviates common challenges in first-person point tracking, such as occlusion or out-of-view, since these conditions also weaken correlation features, which is overlapped by MAS operations.

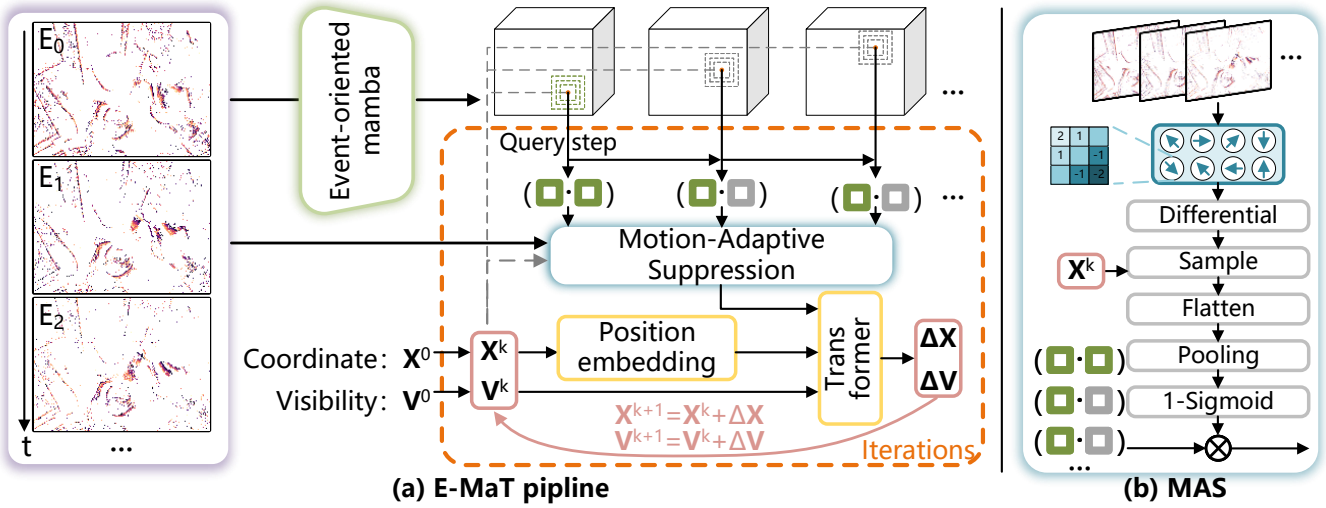


Figure 2: (a) Overview of E-MaT, which consists of an initialization phase and an iterative refinement phase. In the initialization phase, the event-oriented mamba extracts temporally aware spatial features from time surfaces. During iterative refinement, E-MaT samples from the these features based on the query points, computes correlations, and applies (b) the Motion-Adaptive Suppression (MAS) module to suppress unstable time steps with strong motion variations, enhancing robustness.

Depicted in Figure 2(b), MAS first applies 8 directional Sobel operators to the time surface $\mathcal{E}$ to compute motion intensity maps at all time steps, resulting in a directional gradient tensor $\mathcal{M} \in \mathbb{R}^{T\times 8\times H\times W}$. Note that a similar gradient process is used EOM, but MAS is specifically designed to capture changes in motion intensity, as these directly cause fluctuations in event trigger frequencies. In contrast, EOM focuses on extracting the dominant motion direction and local motion intensity. Therefore, MAS further operates on $\mathcal{M}$ using a temporal difference operation, effectively modeling acceleration. Given the query point $\mathbf{X}_k$, MAS samples motion variation at corresponding spatial locations and flattens the sampled values. It then performs global average pooling to summarize directional motion dynamics. The resulting motion variation descriptor is passed through a non-linear suppression function $1-\sigma(\cdot)$, where $\sigma(\cdot)$ denotes the sigmoid function, to compute an inverse confidence weight for each timestep. This weight is then applied to the correlation features $\mathbf{C}^k$ via element-wise multiplication:

$$\hat{\mathbf{C}}^k = (1-\sigma(\mathrm{GAP}(\Delta\mathcal{M}))) \odot \mathbf{C}^k \tag{3}$$

After this, $\hat{\mathbf{C}}^k$, the position-encoded $\mathbf{X}_k$, and visibility $\mathbf{V}_k$ are used together to refine the values of $\Delta\mathbf{X}$ and $\Delta\mathbf{V}$.

Supervision

The total loss consists of two components: regression loss and classification loss. For the regression part, a Huber loss with a delta of 6 is applied at each iteration:

$$L_{reg} = \sum_{i=1}^{K} \gamma^{K-i} Huber(\mathbf{Y}^i, \mathbf{Y}_{gt}). \tag{4}$$

Here $K=4$ represents the total number of iterations, and $\gamma=0.8$ means greater penalties are imposed on later iterations. For the classification part, binary cross entropy loss

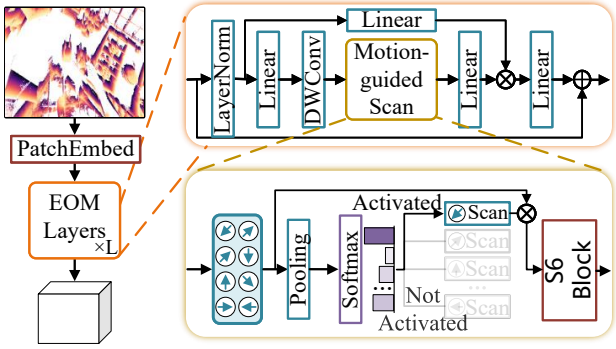


Figure 3: Architecture of the event-oriented Mamba. It adaptively selects the dominant motion direction to determine patch order and modulates features by local motion strength, enabling effective temporal modeling.

is used at each iteration to penalize the difference between visibility and the ground truth:

$$L_{cls} = \sum_{i=1}^{K} \gamma^{K-i} BCE(\mathbf{V}^i, \mathbf{V}_{gt}). \tag{5}$$

The final loss is defined as $L = L_{reg} + L_{cls}$.

Dataset Collection

DVS-EgoPoints Dataset

Due to the lack of event-based datasets for egocentric point tracking, we introduce DVS-EgoPoints, a dual-modality dataset captured using a DAVIS346 camera. It provides synchronized event streams and videos at a resolution of 346×260 and 25 FPS. The dataset targets challenging scenarios such as low light and fast motion, with recordings

Methods	Modality	DVS-EgoPoints						Ev-EgoPoints					
		δ_{avg}	δ^*_{avg}	ReID	IVA	OOVA	OA	δ_{avg}	δ^*_{avg}	ReID	IVA	OOVA	OA
PIPs++	Video	34.9	63.5	12.2	93.4	39.7	-	36.6	58.1	16.8	89.9	52.0	-
CoTracker	Video	34.1	61.9	11.1	94.9	39.2	74.6	39.6	57.5	7.2	78.1	82.0	81.8
CoTracker3	Video	36.1	63.6	15.6	99.8	19.7	79.4	50.5	74.8	14.3	**99.1**	31.4	**85.0**
E2Vid + PIPs++	Event	29.9	55.2	6.7	93.3	41.4	-	-	-	-	-	-	-
E2Vid + CoTracker3	Event	31.1	54.4	2.2	**100.0**	13.9	72.0	-	-	-	-	-	-
ETAP	Event	36.2	64.1	14.9	97.3	21.5	75.1	50.7	72.3	13.6	91.2	33.8	81.5
MATE	Event	35.9	64.6	13.5	89.3	41.5	-	37.5	57.4	16.9	76.5	62.6	-
Ours	Event	**39.8**	**67.8**	**17.1**	91.2	**50.1**	**82.0**	**51.2**	**75.5**	**18.1**	91.6	**82.7**	82.3

Table 1: The performance of the evaluated trackers on the DVS-EgoPoints and Ev-EgoPoints datasets. δ_{avg} and δ^*_{avg} reflect the tracking precision, with higher values indicating better performance. ReID evaluates the re-identification precision for reentering points. IVA, OOVA, and OA assess the accuracy of point visibility classification. All models are retrained on K-EPIC-Points to improve performance in egocentric point tracking. Best results are in **bold**.

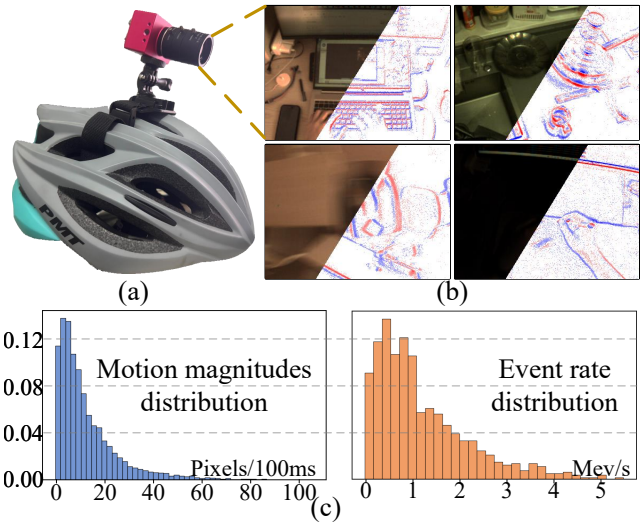


Figure 4: (a) Recording setup with a DAVIS346 mounted on a helmet. (b) Four DVS-EgoPoints samples, each showing a frame at the top left and the corresponding event frame at the bottom right, covering challenges like motion blur (bottom left) and low light (bottom right).(c) Characteristics of the data distribution in DVS-EgoPoints.

captured both day and night. As shown in Figure 4, volunteers wore a helmet-mounted camera while performing unconstrained activities across various indoor scenes (e.g., kitchens, bedrooms, offices). It includes two scenarios of varying difficulty: in the re-identification setting, subjects move away and return, requiring the model to re-establish correspondences under significant viewpoint changes, while in continuous tracking, query points remain in view, such as when gazing at a handheld object. Figure 4 also illustrates samples under fast motion and low light, where video frames may degrade but events remain robust. The dataset exhibits a wide motion distribution, with peak displacement exceeding 100 pixels per 100 ms, and the event rate can reach 5 million events/s, see Figure 4(c).

Annotations keeps the same with (Darkhalil et al. 2024). For each sequence, three frames are manually selected: one reference frame and two evaluation frames. All annotations are performed by a single expert for consistency. Annotators are instructed to label approximately 10 points per reference frame. In evaluation frames, each annotated point is categorized as: (1) **OCC**: within the field of view but occluded; (2) **ReID**: visible in the field of view but previously left it during the sequence. (3) **OOV**: completely out of view. Both OCC and ReID are also marked as in-view (IV).

Simulated Datasets

In this work, events are synthesized from both datasets (Darkhalil et al. 2024) using v2e (Hu, Liu, and Delbruck 2021). Following the original setup, one for training, while the other is referred to as Ev-EgoPoints for evaluation.

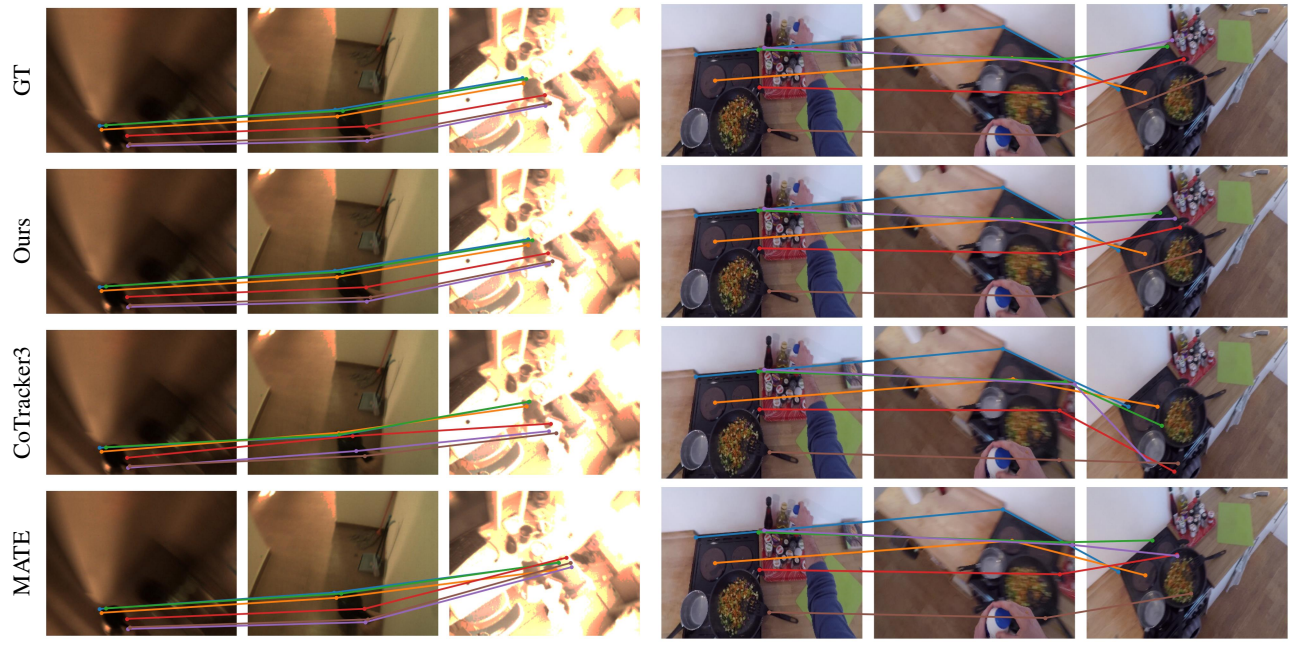
Experiment

Implementation Details

The model is trained on event modality of K-EPIC-Points at a spatial resolution of 384×512. The AdamW optimizer (Loshchilov 2017) is used for optimization, along with the OneCycleLR scheduler (Smith and Topin 2019). The maximum learning rate is set to $5e-4$, and the cycle percentage is configured at 0.05. Experiments are conducted in parallel on 8 Nvidia RTX A6000 GPUs, implemented in PyTorch.

Quantitative Results

Metrics. The evaluation follows the setup of (Darkhalil et al. 2024), including standard point tracking metrics and re-identification indicators. δ_{avg} denotes the percentage of trajectories with average error within $1, 2, 4, 8, 16$, averaged across thresholds. δ^*_{avg} uses a relaxed set $8, 16, 24$ for challenging egocentric data. ReID measures the percentage of trajectories that are successfully re-identified after the query point re-enters the field of view, where success means tracking error within $\{8, 16, 24\}$, averaged across the thresholds. IVA (In-View Accuracy) is the ratio of correctly predicted in-view points ,including visible and occluded; Out-of-View



(a) DVS-EgoPoints result

(b) Ev-EgoPoints result

Figure 5: Qualitative results on DVS-EgoPoints and Ev-EgoPoints datasets. The left is the reference frame; the middle and right are evaluation frames. Dots represent query points, with lines connecting matching points across frames. For clarity, event-based results are mapped onto synchronized frames for direct comparison. E-MaT ensures stable tracking by leveraging events' high temporal resolution, even under motion blur or overexposure. (a) DVS-EgoPoints sample, with an interval of 314 frames, lasting about 13s. (b) Ev-EgoPoints sample, with an interval of 994 frames, lasting about 20s.

Accuracy (OOVA) measures accuracy for out-of-view predictions; Occlusion Accuracy (OA) is for invisible points.

Results. Table 1 compares E-MaT against video-based and event-based trackers. Additionally, to ensure a fair comparison, event-to-video reconstruction pipelines are also included, in which event data is first converted into video using E2Vid (Rebecq et al. 2019) and processed by video-based trackers. Results on Ev-EgoPoints omit reconstruction-based baselines, as the event modality in this dataset is synthesized from video, making reconstruction redundant. Moreover, OA is not reported for PIPs++ and MATE due to the lack of visibility modeling in the original methods.

The proposed method outperforms existing approaches across most metrics on both datasets. Video-based trackers struggle with motion blur under rapid motion, limiting localization accuracy. Reconstruction-based pipelines compromise event fidelity due to inconsistencies introduced during video conversion. Existing third-person event-based methods follow a similar pipeline to video-based ones—extracting spatial features before temporal modeling—causing loss of fine-grained temporal cues. E-MaT addresses these issues via motion-guided temporal modeling that preserves the spatiotemporal dynamics of events. Notably, CoTracker3 achieves high IVA by predicting most query points as in-view, even after they leave the field of view—resulting in inflated IVA but low OOVA.

Performance gains are evident on DVS-EgoPoints, which features low lighting and severe motion blur, highlighting the value of events for egocentric point tracking.

Scan direction				EOM	MAS	DVS-EgoPoints	
H	V	D	Anti-D			δ_{avg}	δ^*_{avg}
✓						37.2	66.4
	✓					36.2	63.9
		✓				35.7	64.2
			✓			35.8	64.5
✓	✓					36.0	64.6
✓	✓	✓	✓			36.3	65.0
				✓		38.6	67.3
				✓	✓	**39.8**	**67.8**

Table 2: Ablation of E-MaT components. "H": Horizontal. "V": Vertical. "D": Diagonal. "Anti-D": Anti-diagonal

Qualitative Analysis

As shown in Figure 5, E-MaT demonstrates advantages over video-based and event-based trackers on both datasets. On DVS-EgoPoints, CoTracker3 suffers from degraded imaging quality under challenging lighting conditions, leading to failures in appearance-based matching. Moreover, its limited temporal resolution hinders the capture of rapid motion caused by egocentric viewpoint shifts. Although MATE leverages high-temporal-resolution events as input, it follows a video-like paradigm that models spatial and temporal components in a decoupled manner, resulting in the loss of

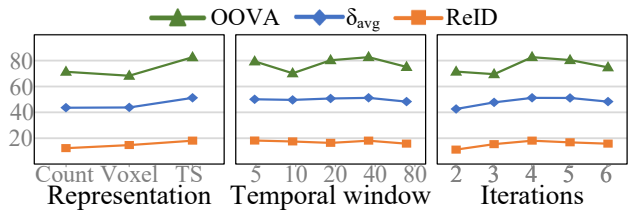
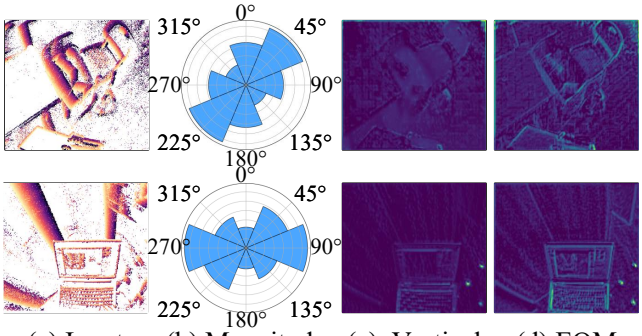


Figure 6: Impact of different input settings.



(a) Input (b) Magnitude (c) Vertical (d) EOM

Figure 7: (a) Input time surface; (b) Motion magnitudes across directions—note that opposing directions yield equal absolute values. (c) Vertical scanning leads to diffuse responses due to disrupted temporal order; (d) EOM enhances activations along the dominant motion path, producing sharper, temporally-aware features.

fine-grained temporal information and causing point drift. In contrast, E-MaT encodes spatial features along the temporal motion trajectory, enabling more reliable feature association and enhancing tracking stability. On Ev-EgoPoints, although lighting is generally favorable, sequences involve frequent transitions between fast and slow motion. MATE and CoTracker3 accumulate errors during these transitions, causing gradual tracking accuracy degradation.

Ablation Study

Event representation: Figure 6 compares event count, voxel grid, and time surface. The time surface yields the best performance by preserving local motion history, which is essential for robustness under motion blur and occlusion.

Temporal window: The Temporal window size of time surface affects how much motion is encoded. As shown in Figure 6, a 40ms window performs best. Shorter windows fail to capture meaningful motion patterns, while longer ones introduce noise from outdated events.

Number of Iterations: Figure 6 shows that four iterations provide the best balance. Fewer iterations underfit the refinement process while more iterations cause overfitting.

EOM: Different scan strategies are evaluated in Table 2: single-direction, multi-directional (e.g., H&V), and EOM. Adding more directions does not improve performance; instead, it disrupts the temporal structure of the time surface, reducing feature discriminability.

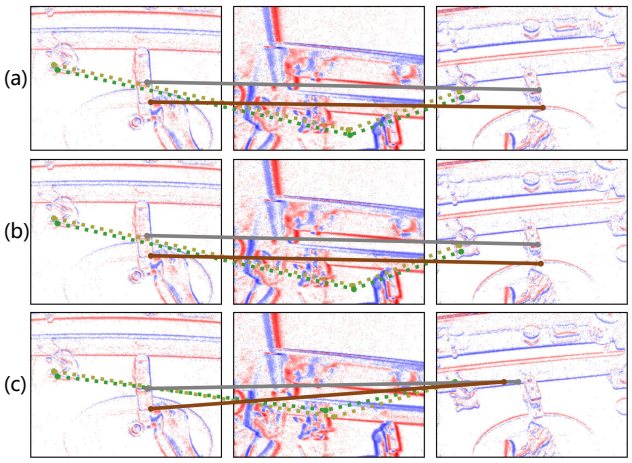


Figure 8: Effectiveness of MAS. The left column shows the reference frame; the middle and right are evaluation frames. Dashed lines indicate points that are occluded during tracking, and solid lines indicate points that move out of the field of view during tracking. (a) Ground truth. (b) Result with MAS. (c) Result without MAS.

MAS: MAS suppresses correlation responses at time steps with unstable motion, encouraging reliance on consistent temporal cues. This helps mitigate occlusion and out-of-view errors, as illustrated in Figure 8. Adding MAS leads to a 1.2 improvement on δ_{avg} according to Table 2.

Conclusion

This paper presents E-MaT, a novel event-based egocentric point tracking framework that introduces an event-oriented mamba encoder. The encoder constructs feature modeling paths aligned with dominant motion trend extracted from events and modulates feature propagation based on local motion intensity, enabling stable spatialtemporal representation. In addition, a motion-adaptive suppression module improves temporal robustness by suppressing unreliable correlation features according to motion intensity variations, mitigating the effects of abrupt motion and partial observability. To advance research in this domain, a real-world egocentric point tracking dataset, DVS-EgoPoints, incorporating both event and video modalities and covering challenging scenarios such as low light and fast motion, is introduced. Experiments on both DVS-EgoPoints and a synthetic benchmark demonstrate that E-MaT outperforms state-of-the-art methods, showing significant potential for applications in embodied intelligence, AR/VR, and related fields.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207, 62306295, 62576328, and 62436008. The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

References

Almatrafi, M.; Baldwin, R.; Aizawa, K.; and Hirakawa, K. 2020. Distance surface for event-based optical flow. *IEEE transactions on pattern analysis and machine intelligence*, 42(7): 1547–1556.

Behrens, T.; Zurbrügg, R.; Pollefeys, M.; Bauer, Z.; and Blum, H. 2025. Lost & Found: Tracking Changes From Egocentric Observations in 3D Dynamic Scene Graphs. *IEEE Robotics and Automation Letters*, 10(4): 3739–3746.

Cho, S.; Huang, J.; Nam, J.; An, H.; Kim, S.; and Lee, J.-Y. 2024. Local All-Pair Correspondence for Point Tracking.

Darkhalil, A.; Guerrier, R.; Harley, A. W.; and Damen, D. 2024. EgoPoints: Advancing Point Tracking for Egocentric Videos.

Doersch, C.; Gupta, A.; Markeeva, L.; Recasens, A.; Smaira, L.; Aytar, Y.; Carreira, J.; Zisserman, A.; and Yang, Y. 2022. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; and Scaramuzza, D. 2022. Event-Based Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 154–180.

Gao, Y.; Lu, J.; Li, S.; Ma, N.; Du, S.; Li, Y.; and Dai, Q. 2023. Action recognition and benchmark using event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 14081–14097.

Ge, C.; Fu, X.; He, P.; Wang, K.; Cao, C.; and Zha, Z.-J. 2025. EventMamba: Enhancing Spatio-Temporal Locality with State Space Models for Event-Based Video Reconstruction. *arXiv preprint arXiv:2503.19721*.

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

Hamann, F.; Gehrig, D.; Febryanto, F.; Daniilidis, K.; and Gallego, G. 2025. ETAP: Event-based Tracking of Any Point. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27186–27196.

Han, H.; Zhai, W.; Cao, Y.; Li, B.; and jun Zha, Z. 2025. MATE: Motion-Augmented Temporal Consistency for Event-based Point Tracking.

Harley, A. W.; Fang, Z.; and Fragkiadaki, K. 2022. Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. In *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, 59–75. Springer Nature Switzerland.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, Y.; Liu, S.-C.; and Delbruck, T. 2021. v2e: From video frames to realistic DVS events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1312–1321.

Huang, J.; Wang, S.; Wang, S.; Wu, Z.; Wang, X.; and Jiang, B. 2024. Mamba-fetrack: Frame-event tracking via state space model. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 3–18. Springer.

Karaev, N.; Makarov, I.; Wang, J.; Neverova, N.; Vedaldi, A.; and Rupprecht, C. 2024. CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos.

Karaev, N.; Rocco, I.; Graham, B.; Neverova, N.; Vedaldi, A.; and Rupprecht, C. 2023. CoTracker: It Is Better to Track Together.

Kim, I. H.; Cho, S.; Huang, J.; Yi, J.; Lee, J.-Y.; and Kim, S. 2025. Exploring Temporally-Aware Features for Point Tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1962–1972.

Li, H.; Zhang, H.; Liu, S.; Zeng, Z.; Ren, T.; Li, F.; and Zhang, L. 2025. TAPTR: Tracking Any Point with Transformers as Detection. In *Computer Vision – ECCV 2024*, 57–75. Springer Nature Switzerland.

Li, S.; Singh, H.; and Grover, A. 2025. Mamba-ND: Selective State Space Modeling for Multi-Dimensional Data. In *Computer Vision – ECCV 2024*, 75–92. Springer Nature Switzerland.

Liao, B.; Zhai, W.; Wan, Z.; Cheng, Z.; Yang, W.; Zhang, T.; Cao, Y.; and Zha, Z.-J. 2024. Ef-3dgs: Event-aided free-trajectory 3d gaussian splatting. *arXiv preprint arXiv:2410.15392*.

Lichtsteiner, P.; Posch, C.; and Delbruck, T. 2008. A 128 $\times$ 128 120 dB 15 μ s latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2): 566–576.

Liu, J.; Wang, B.; Tan, Z.; Zhang, J.; Shen, H.; and Hu, D. 2024a. Tracking any point with frame-event fusion network at high frame rate.

Liu, M.; and Delbruck, T. 2018. Adaptive Time-Slice Block-Matching Optical Flow Algorithm for Dynamic Vision Sensors. In *British Machine Vision Conference (BMVC)*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024b. VMamba: Visual State Space Model. *Advances in Neural Information Processing Systems*, 37: 103031–103063.

Loshchilov, I. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mueggler, E.; Bartolozzi, C.; and Scaramuzza, D. 2017. Fast Event-based Corner Detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, 33–1.

Plizzari, C.; Goletto, G.; Furnari, A.; Bansal, S.; Ragusa, F.; Farinella, G. M.; Damen, D.; and Tommasi, T. 2024. An outlook into the future of egocentric vision. *International Journal of Computer Vision*, 132(11): 4880–4936.

Qu, J.; Li, H.; Liu, S.; Ren, T.; Zeng, Z.; and Zhang, L. 2024. TAPTRv3: Spatial and Temporal Context Foster Robust Tracking of Any Point in Long Video.

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.

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): 7742–7759.

Smith, L. N.; and Topin, N. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, 369–386. SPIE.

Tan, G.; Wang, Y.; Han, H.; Cao, Y.; Wu, F.; and Zha, Z.-J. 2022. Multi-Grained Spatio-Temporal Features Perceived Network for Event-Based Lip-Reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20094–20103.

Taverni, G.; Moeys, D. P.; Li, C.; Cavaco, C.; Motsnyi, V.; Bello, D. S. S.; and Delbruck, T. 2018. Front and back illuminated dynamic and active pixel vision sensors comparison. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(5): 677–681.

Wan, Z.; Wang, Y.; Tan, G.; Cao, Y.; and Zha, Z.-J. 2022. S2N: Suppression-Strengthen Network for Event-Based Recognition Under Variant Illuminations. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 716–733. Springer Nature Switzerland.

Wan, Z.; Wang, Y.; Wei, Z.; Tan, G.; Cao, Y.; and Zha, Z.-J. 2024. Event-based optical flow via transforming into motion-dependent view. *IEEE Transactions on Image Processing*.

Wan, Z.; Zhai, W.; Cao, Y.; and Zha, Z. 2025. EMoTive: Event-guided Trajectory Modeling for 3D Motion Estimation. *arXiv preprint arXiv:2503.11371*.

Wang, X.; Wang, S.; Wang, X.; Zhao, Z.; Zhu, L.; Jiang, B.; et al. 2024. Mambaevt: Event stream based visual object tracking using state space model. *arXiv preprint arXiv:2408.10487*.

Zhao, A.; Tang, C.; Wang, L.; Li, Y.; Dave, M.; Tao, L.; Twigg, C. D.; and Wang, R. Y. 2025. EgoBody3M: Egocentric Body Tracking on a VR Headset Using a Diverse Dataset. In *Computer Vision – ECCV 2024*, 375–392. Springer Nature Switzerland.

Zheng, Y.; Harley, A. W.; Shen, B.; Wetzstein, G.; and Guibas, L. J. 2023. PointOdyssey: A Large-Scale Synthetic Dataset for Long-Term Point Tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19855–19865.

Zhu, A. Z.; Yuan, L.; Chaney, K.; and Daniilidis, K. 2019. Unsupervised Event-Based Learning of Optical Flow, Depth, and Egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 989–997.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model.

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.