

# Maximizing Asynchronicity in Event-based Neural Networks

Haiqing Hao<sup>1</sup>, Nikola Zubić<sup>2</sup>, Weihua He<sup>1</sup>, Zhipeng Sui<sup>1</sup>, Davide Scaramuzza<sup>2</sup>, and Wenhui Wang<sup>1</sup>

<sup>1</sup>Department of Precision Instrument, Tsinghua University

<sup>2</sup>Robotics and Perception Group, University of Zurich

## Abstract

Event cameras deliver visual data with high temporal resolution, low latency, and minimal redundancy, yet their asynchronous, sparse sequential nature challenges standard tensor-based machine learning (ML). While the recent *asynchronous-to-synchronous* (A2S) paradigm aims to bridge this gap by asynchronously encoding events into learned representations for ML pipelines, existing A2S approaches often sacrifice representation expressivity and generalizability compared to dense, synchronous methods. This paper introduces **EVA**(**E**vent **A**synchronous representation learning), a novel A2S framework to generate highly expressive and generalizable event-by-event representations. Inspired by the analogy between events and language, EVA uniquely adapts advances from language modeling in linear attention and self-supervised learning for its construction. In demonstration, EVA outperforms prior A2S methods on recognition tasks (DVS128-Gesture and N-Cars), and represents the first A2S framework to successfully master demanding detection tasks, achieving a remarkable 47.7 mAP on the Gen1 dataset. These results underscore EVA’s transformative potential for advancing real-time event-based vision applications.

## 1 Introduction

Traditional frame-based cameras synchronously capture dense visual intensity data at fixed rates. In contrast, event cameras asynchronously detect pixel-level intensity changes, resulting in event sequences characterized by high temporal resolution (up to 1  $\mu$ s) and low spatial redundancy [10]. Due to their ability to capture rapid movements with minimal latency and reduced bandwidth requirements, event cameras have garnered widespread attention in autonomous driving [13], robotics [16], and biomedical analysis [19]. However, the asynchronous and sparse nature of event data inherently complicates its direct and efficient use with standard machine learning (ML) algorithms, which typically demand tensor-based inputs. Current approaches to utilize event data in vision tasks predominantly fall into two categories: converting event sequences into dense, image-like representations aggregated over fixed time windows [42, 26, 12, 20, 50], or employing spiking neural networks (SNNs) [19]. The former often fails to fully exploit the inherent sparsity and rich temporal information, while the latter relies on specialized hardware and encounters challenges in training high-performance models effectively [30].

To better leverage the fine-grained temporal feature and inherent sparsity of events while benefiting from advancements in standard ML, the *asynchronous-to-synchronous* (A2S) paradigm has recently emerged [30]. This paradigm focuses on designing an efficient asynchronous encoder that recurrently

Emails: hhq23@mails.tsinghua.edu.cn, zubic@ifi.uzh.ch, hwh20@mails.tsinghua.edu.cn, hzp22@mails.tsinghua.edu.cn, sdavide@ifi.uzh.ch, ww@tsinghua.edu.cn

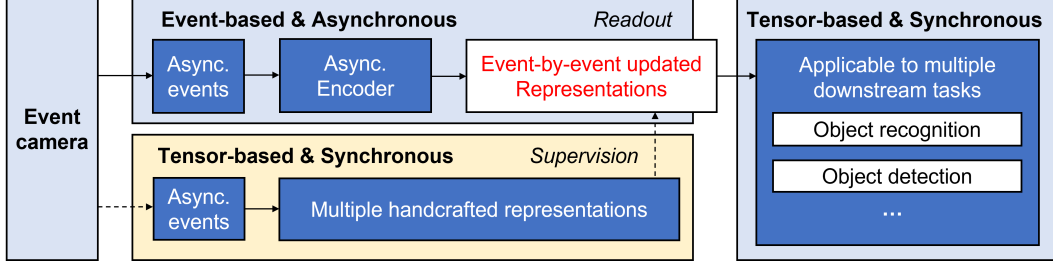


Figure 1: Overview of the proposed EVA A2S framework. It features an asynchronous LA-based encoder (top middle) producing event-by-event updated representations for synchronous downstream tasks (right). These representations are learned via a self-supervised module (bottom middle) and are event-by-event updated during inference to be sampled on-demand by downstream tasks.

integrates events into a tensor-based representation on an event-by-event basis in real time. This representation can then be synchronously sampled at flexible, on-demand rates by downstream ML tasks. However, existing A2S methodologies [30] rely on preliminary models with limited expressivity as a compromise for computational efficiency. Consequently, they achieve suboptimal performance on complex tasks compared to dense, tensor-based approaches [14, 12, 51]. Furthermore, these representations are typically learned in a supervised manner, limiting their generalizability across diverse downstream applications.

To design an efficient A2S framework with more expressivity and generalizability, we take inspiration from the analogy between events and language [43]. On one hand, events and language are both organized in sequences. Events record incremental visual changes, similar to words in a sentence, incrementally contributing to the semantic meaning. This inspires us to treat each event as a discrete token and leverage recent advances in natural language processing (NLP), including linear attention (LA) for efficient sequence modeling [22, 47, 32, 45] and auto-regressive prediction [38] for self-supervised learning. On the other hand, crucial distinctions between events and language exist, primarily in information density and spatial locality. A single language token inherently possesses significant semantic meaning, whereas an individual event offers limited information, gaining significance primarily through aggregation within its historical context. This distinction steers our focus towards the aggregated global information rather than isolated event instances. Additionally, events typically signify spatially localized changes, contrasting with language, which lacks inherent spatial attributes. This motivates modeling correlations primarily between spatially adjacent events to reduce complexity.

In this work, we introduce a new A2S framework with a more expressive asynchronous encoder, which achieves event-by-event representation updating, and a self-supervised method for highly generalizable representation learning. Our asynchronous encoder is built on the basis of LA, which enables recurrent inference and parallel training. We propose to use the matrix-value hidden state (MVHS) introduced in LA as output to expand the model’s expressivity on the aggregated global information and propose patch-wise encoding to leverage the locality. For ease of implementation and better performance, our model is built on a high-performance open source LA architecture RWKV-6 [33], which implements MVHS and is stable to train. We propose a self-supervised method for generalizable representation learning, which consists of two tasks: multi-representation prediction (MRP) and next-representation prediction (NRP). The former encourages the representation to learn diverse information from multiple handcrafted (converted) representations [26, 42] comprehensively, and the latter, motivated by next-token prediction strategies prevalent in NLP [38], pushes the model beyond simple memorization. Our framework is superior to the previous A2S work [30] on recognition tasks, and it is the first A2S framework, to the best of our knowledge, that successfully masters demanding detection tasks, achieving a remarkable 47.7 mAP on the Gen1 dataset. We name our framework as **E**Vent **A**synchronous representation learning (EVA). Our contributions include:

1. An asynchronous LA-based encoder architecture derived from RWKV-6, enabling efficient event-by-event representation updating with improved expressivity.
2. A novel multi-task SSL method that learns generalizable representations across various downstream tasks.

3. We present EVA, a new A2S framework that synergizes the asynchronous encoder with self-supervised representation learning. Compared to the previous A2S work, EVA outperforms on recognition tasks and, for the first time, successfully masters demanding detection tasks with a remarkable 47.7 mAP on the Gen1 dataset.

## 2 Related Work

### 2.1 Asynchronous event processing

To fully leverage the sparsity and high temporal resolution of event data within deep learning, several recent efforts have focused on asynchronous event processing. These include point cloud approaches [41, 30], convolutional methods [31, 4], graph-based models [28, 40, 52], attention-based mechanisms [21], and LSTM-based solutions [5, 39]. Specifically, EventNet [41] introduces an event-by-event feature update mechanism based on PointNet [36]. However, features are encoded into a vector and therefore limited in expressivity. Convolutional methods have explored deriving local update rules of activations in standard [4] or sparse [31] convolution layers. While they can leverage sparsity and reduce redundant computation, they often discard the temporal component of events. Graph-based methods [28, 40, 52] transform spatio-temporal events into graphs and introduce local update rules for event-by-event processing, thereby capturing both spatial and temporal information. However, graph-based methods are known to suffer from over-smoothing, which limits their scalability and performance on intricate patterns. An attention-based, memory-augmented method was proposed in [21] to process batched events at each timestep. The LSTM-based Matrix-LSTM [5] method learns and recurrently updates a representation with a pixel-wise LSTM for downstream tasks. FARSE-CNN [39] integrates LSTMs within the neurons of convolutional layers to capture spatio-temporal information and employs local update rules. Our work is most related to the A2S method ALERT-Transformer [30]. In [30], a patch-wise EventNet is used for asynchronous feature encoding, and the feature is forwarded to a synchronous Transformer for downstream tasks. Our method follows the same A2S paradigm but introduces a new architecture with improved expressivity, solving the problem of limited expressivity brought up by the EventNet encoder.

### 2.2 Event data-based self-supervised representation learning

Event camera data-based self-supervised representation learning can be broadly categorized into methods that learn only from events and those that incorporate other modalities. For event-only pre-training, MEM [25] uses the masked image modeling method [2] to event histograms for pretraining. Other works [3, 49] utilize contrastive pretraining with joint embedding architectures [7, 17] on event histograms. These methods typically treat event histograms as images and adapt established image pretraining strategies to the event camera domain. When learning with other modalities, event data are often paired with conventional intensity images for representation learning. In [48], contrastive pretraining is explored on event data synthetically converted from RGB images and their source images. EVRepSL [37] learns representations by exploiting the relationship between intensity images and events recorded simultaneously by ATIS cameras. These approaches rely on the correlations between paired visual modalities to guide representation learning.

## 3 Method

Our framework’s design leverages the event-language analogy. We first analyze their commonalities and distinctions. On one hand, events and language are similar in their (i) **sequential nature** and (ii) **incremental manner** of context contribution. Events convey local intensity changes, similar to language tokens that incrementally build the context (Figure 2). On the other hand, key distinctions exist in their (i) **information density** and (ii) **spatial locality**. An individual language token has explicit semantics. In contrast, events only record pixel-level intensity changes and require temporal aggregation to become informative.

As shown in Figure 3, our A2S framework first transforms events into tokens and embeds them with their spatial and temporal attributes. These embeddings are then asynchronously encoded into representations. The representations are learned via self-supervision during training and are generated event-by-event at inference for real-time downstream applications.

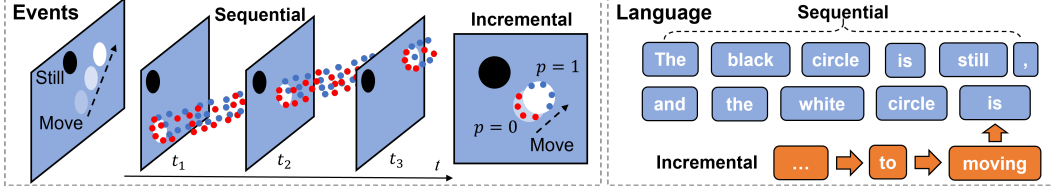


Figure 2: Parallels between event data and language: their sequential nature and incremental manner.

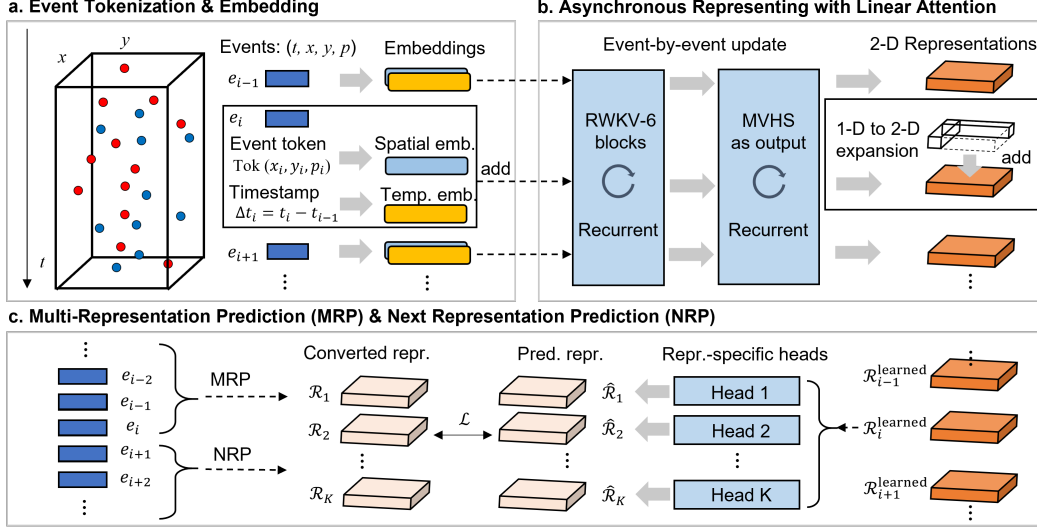


Figure 3: Architecture of the A2S framework: (a) Tokenizing and embedding events. (b) Asynchronous encoding via LA. (c) MRP and NRP for self-supervised representation learning.

### 3.1 Architecture: asynchronous encoder

Key challenges in designing an asynchronous event encoder encompass: (i) allowing recurrent inference for real-time representation updates and fast parallel training on long event sequences, (ii) continuously encoding events with variable temporal intervals, and (iii) efficiently and expressively aggregating information into the representation. To address the first challenge, drawing from the event-language analogy, we find that recent linear attention (LA) in sequential language modeling supports both recurrent inference and parallel training. Building on LA, we incorporate temporal embedding, MVHS as output, and patch-wise encoding to resolve the other two challenges.

#### 3.1.1 Event tokenization

We define an event sequence as  $\mathcal{E} = \{e_i = (t_i, x_i, y_i, p_i), i = 1, 2, \dots\}$ , where  $t, x, y, p \in \mathbb{R}$  signify timestamp, coordinates, and polarity of events, with  $p \in \{0, 1\}$  indicating the two polarities. Each event's spatial component is tokenized using a bijection Tok as:

$$\begin{aligned} \text{Tok} : \mathcal{E} &\rightarrow \mathbb{N} \\ (x, y, p) &\mapsto p \times H \times W + y \times W + x, \end{aligned} \quad (1)$$

where  $H, W$  denote the height and width of the spatial area containing the event sequence. The vocabulary size is thus  $2 \times H \times W$ . This mapped token signifies the event's spatial location.

#### 3.1.2 Embedding events with varied temporal intervals

Each tokenized event  $e_i = (t_i, \text{Tok}(x_i, y_i, p_i))$  is then embedded as an input vector  $x_i \in \mathbb{R}^D$  to the encoder. For event tokens that represent spatial locations, we employ a learnable linear embedding on the tokens and get  $\text{Emb}_{\text{spatial}}(e_i) \in \mathbb{R}^D$ . Events are inherently asynchronous, so temporal information must be embedded in the input. Considering that event cameras generate events at high temporal

resolution, the absolute timestamp (in  $\mu s$ ) quickly grows large in continuous operation; embedding the absolute timestamp would push the model beyond the range observed during the pretraining period, leading to the same kind of length-extrapolation failures seen in language models [35, 44, 6]. We therefore embed the time difference  $\Delta t_i = t_i - t_{i-1}$  instead of the absolute timestamp, using the sinusoidal embedding [46] on the  $\Delta t_i$  as:

$$\text{Emb}_{\text{temporal}}(e_i)_k = \begin{cases} \sin(\Delta t_i / 10000^{\frac{2k}{D}}), & \text{if } k \text{ even,} \\ \cos(\Delta t_i / 10000^{\frac{2k}{D}}), & \text{if } k \text{ odd,} \end{cases} \quad k = 0, 1, \dots, D-1. \quad (2)$$

The final embedding is then formed by summing the spatial and temporal embeddings:

$$\text{Emb}(e_i) = \text{Emb}_{\text{spatial}}(e_i) + \text{Emb}_{\text{temporal}}(e_i) \in \mathbb{R}^D. \quad (3)$$

### 3.1.3 Modeling event sequence with linear attention

**Background: Linear attention in RWKV-6.** The asynchronous encoder is based on RWKV-6, a high-performance linear attention architecture for language modeling[33]. For an input sequence  $\{\mathbf{x}_i \in \mathbb{R}^{D \times 1}\}_{i=1}^T$  of length  $T$ , a single RWKV-6 block performs token mixing (TM) and channel mixing (CM).

In TM, inputs are first mapped to  $\{\mathbf{r}_i, \mathbf{w}_i, \mathbf{k}_i, \mathbf{v}_i \in \mathbb{R}^{D \times 1}\}_{i=1}^T$ . and subsequently mixed by the *rwkv* operator:

$$\mathbf{y}_i = \left[ \sum_{t=1}^{i-1} \text{diag} \left( \bigotimes_{j=t+1}^{i-1} \mathbf{w}_j \right) \mathbf{k}_t \mathbf{v}_t^T + \text{diag}(\mathbf{u}) \mathbf{k}_i \mathbf{v}_i^T \right] \mathbf{r}_i \in \mathbb{R}^{D \times 1}, \quad (4)$$

where  $\mathbf{u} \in \mathbb{R}^{D \times 1}$  is a learnable parameter. This *rwkv* operator, involving only element-wise prefix sums and products, enables fast training via parallel scan [29]. Equation 4 can be expressed in a recurrent form as:

$$\begin{aligned} \mathbf{S}_i &= \text{diag}(\mathbf{w}_i) \mathbf{S}_{i-1} + \mathbf{k}_i \mathbf{v}_i^T \in \mathbb{R}^{D \times D}, \\ \mathbf{y}_i &= (\mathbf{S}_{i-1} + \text{diag}(\mathbf{u}) \mathbf{k}_i \mathbf{v}_i^T) \mathbf{r}_i \in \mathbb{R}^{D \times 1}. \end{aligned} \quad (5)$$

Consequently, the hidden state  $\mathbf{S}$  can be updated recursively at inference upon the arrival of each new event token, allowing efficient training and recurrent inference on long, continuous event sequences.

To better capture diverse information aspects, RWKV-6 uses multi-head TM. Each input is split into  $N$  heads of dimension  $D_{\text{head}} = D/N$ , and the *rwkv* operator acts on each head individually. The hidden state in multi-head TM is  $\mathbf{S} \in \mathbb{R}^{N \times D_{\text{head}} \times D_{\text{head}}}$ .

The CM layer transforms TM outputs using a gated feed-forward network, similar to that in Transformers [46]. Appendix A provides details.

**Representing and training with matrix value hidden states (MVHS) as output.** Typical NLP tasks use sequence-to-sequence models mapping 1-D token embeddings to 1-D embeddings (or classifying tokens). We propose direct training on the 2-D MVHS, rather than the 1-D output, and use the MVHS as the event representation. Appendix A provides details of the MVHS layer. This practice offers several benefits. First, MVHS, as the hidden state, naturally contains aggregated information. This aggregated representation is precisely what we want to learn, instead of the mapping from event embeddings to event embeddings in standard auto-regressive pretraining (our attempt to predict the next single event failed to converge). Secondly, MVHS yields an expanded memory of size  $N \times D_{\text{head}} \times D_{\text{head}}$  without increasing model width  $D$ , thus improving the representation expressivity and enabling a lightweight architecture for real-time event processing. Using MVHS as output can reduce model size by approximately  $D_{\text{model}}/N$  versus 1-D outputs (Appendix F). Moreover, the 2-D structure of MVHS helps to learn fine-grained spatial representations, improving performance over prior 1-D features [41, 30].

### 3.1.4 Patch-wise representation encoding

To exploit the locality of the event, we follow the practice in [29] and use patch-wise representation encoding. For an event camera resolution  $(H_{\text{sensor}}, W_{\text{sensor}})$  and patch size  $P$ , we divide events into  $H_{\text{sensor}} \times W_{\text{sensor}}/P^2$  sequences by coordinates. As Figure 4 illustrates, we learn and encode representations for each patch separately. With patch-wise encoding, the model size of the encoder can

be reduced by a factor of approximately  $\# \text{ patches}$ , which helps real-time inference (see Appendix F for discussion). Furthermore, as the asynchronous encoder trains on fixed-size  $P$  patches, this patch-wise encoding allows application to event cameras with differing resolutions [51].

### 3.2 Self-supervised representation learning on asynchronous events

The available A2S method learns the representation in a fully supervised manner [30], leading to limited generalizability to diverse downstream tasks. Here, we propose a self-supervised representation learning method composed of multi-representation prediction (MRP) and next representation prediction (NRP).

#### 3.2.1 Multi-representation prediction

For MRP, we force the learned representation to predict multiple handcrafted (converted) event representations [26, 42, 20], such as event count (EC) and time surface (TS), to learn a more comprehensive representation. For events  $\mathcal{E} = \{e_i = (t_i, x_i, y_i, p_i)\}, i = 1, 2, \dots$ , we expect the asynchronous encoder  $\mathcal{M}_\theta$  parametrized by  $\theta$  to learn an event-by-event representation as  $\mathcal{R}_i^\theta = \mathcal{M}_\theta(\{e_j\}_{j \leq i})$ . Given  $K$  handcrafted representations  $\{\mathcal{R}^k\}_{k=1}^K$ , we predict these representations from  $\mathcal{R}^\theta$  with specific neural network heads parameterized by  $\Theta = \{\theta_k\}_{k=1}^K$ . The objective is:

$$\arg \max_{\theta, \Theta} \mathbb{E}_i \prod_{k=1}^K p(\mathcal{R}^k(\{e_j\}_{j \leq i}) | \mathcal{M}_\theta(\{e_j\}_{j \leq i}), \theta_k). \quad (6)$$

MRP is motivated by the insight that different handcrafted representations capture distinct raw event information aspects and provide visual data for various tasks [50]. We anticipate that learning to predict multiple converted representations will yield a representation generalizable to diverse downstream tasks.

#### 3.2.2 Next representation prediction

Inspired by auto-regressive pretraining of language models, where models learn by next-token prediction [38], NRP forces the encoder to predict handcrafted representations for a future time interval. For a future time window  $\Delta T$ , the objective of NRP is:

$$\arg \max_{\theta, \Theta'} \mathbb{E}_i \prod_{k=1}^{K'} p(\mathcal{R}^k(\{e|t_i < t(e) \leq t_i + \Delta T\}) | \mathcal{M}_\theta(\{e_j\}_{j \leq i}), \theta'_k). \quad (7)$$

We aim for the model to understand inherent motion patterns and beyond just memorize the history by learning to predict the future. In particular, individual events are ineffective prediction targets due to their limited information and unpredictable noise. In contrast, converted representations are more informative and noise-robust, offering a more reliable supervisory signal.

## 4 Experiments

### 4.1 Object recognition

**Datasets.** To compare with previous A2S work [30], we evaluate our work EVA on two event-based recognition tasks: action recognition (DVS128-Gesture [1]) and binary classification (N-Cars [42]). DVS128-Gesture provides 11 classes of human actions recorded by the DVS128 event camera, containing 1342 files of around 6 seconds from 29 subjects under 3 different lighting conditions.

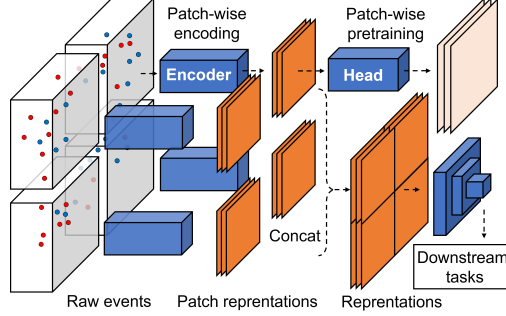


Figure 4: Patch-wise representation encoding. Events are partitioned and encoded by patch. Patch representations could be concatenated for downstream tasks.

N-Cars provides a benchmark for the binary classification task of car recognition, containing 24029 files of 100 ms recorded by the ATIS event camera.

**Implementations.** We present an asynchronous encoder composed of an event embedding with  $D = 128$ , 3 RWKV-6 blocks of width  $D = 128$ , and one MVHS layer as output to generate the learned representation. We use patch size  $P = 16$ . On DVS128-Gesture, to reduce the latency of downstream classification, we use  $D_{\text{head}} = 8$  for a  $2\times$  shrink in height and width of the representation and use a light-weighted 14 layer ResNet [18] (referred to as ResNet-14) for classification. On N-Cars however, since the samples are already cropped small [42], we also present an encoder with  $D_{\text{head}} = 16$  for a full height and weight representation. To maintain the representation size, we use only half-channels of the MVHS output. The model implementation details are given in Appendix C.

On DVS128-Gesture, the encoder is pretrained on event sequences of length 8192 with batch size 32 and Adam optimizer [24] with learning rate 0.001. It takes around 10 hours to train 100 epochs with an RTX 3090 GPU. We use EC and TS for MRP and EC for NRP as SSL targets. Calculating the loss of the predictions and targets at every individual event would demand prohibitive GPU memory. To mitigate this, we divide the event sequence into chunks of length  $T_{\text{chunk}} = 512$ , and calculate the loss only at the end of each chunk. On N-Cars, the encoder is pretrained with a sequence length of 512 and chunk length 16 with batch size 32 and learning rate 0.001. It takes 1 hour to train 20 epochs on an RTX 3090 GPU. We use EC and TS for MRP and EC for NRP as well. To balance the multi-task losses, we use uncertainty weights [23]. The details of SSL objectives are given in Appendix D.

**Metrics.** On DVS128-Gesture, following previous work [30], we evaluate the model on samples of length 8192 and report the accuracy on samples (sample accuracy, SA) and the accuracy after majority voting on all the samples of each file (file voting accuracy, FVA). We also report the complexity (MAC) and inference latency on the RTX 3090 GPU. On N-Cars, we report accuracy.

**Results.** On DVS128-Gesture, our model achieves 96.9% FVA and 92.9% SA (see Table 1), outperforming the best result in previous A2S work (94.1% FVA and 84.6% SA) by 2.8% and 8.3% respectively. Compared with ALERT-Tr. (+RM), our EVA achieves better performance with smaller parameter size and complexity. ALERT-Tr. (+LMM) indeed has fewer parameters, but with a much lower accuracy of 72.6% in SA, 20% less accuracy than ours. It is worth noting that our EVA, with a lightweight ResNet classifier, has much lower inference time (1.5 ms) compared with both the LMM and RM. Our asynchronous encoder has a larger inference latency compared with previous work because of its complicated architecture. However, it is smaller than the duration time of an 8192-length sample, which is about 100 ms in DVS128-Gesture. On N-Cars, previous A2S work [30] did not report the results with their larger model (RM), only gave the accuracy of LMM to show its low complexity. Therefore, here we also compare our EVA with (i) synchronous tensor-based methods[34, 25] and (ii) methods that learns the representation from raw events[12, 5]. In comparison with ALERT-Tr. (+LMM), our method has better accuracy and less classifier latency. Since the files in N-Cars are small in number of events and the pretraining on this dataset is not sufficient, we also present the result with representations generated by an encoder pretrained on Gen1 for detection tasks (EVA-L in Table 3). For a fair comparison with [12] and [5], we follow their practice and use ResNet-34 pretrained on ImageNet[9] as the classifier. Our model, with the encoder pretrained on Gen1 and ResNet-34, has an accuracy of 96.3%, higher than the methods with learned representations.

Table 1: Results on DVS128-Gesture. Latency refers to the time to process a sample of length 8192.

Model	# Params	MAC per		Latency	Acc.	
		Event	Sample		SA	FVA
ALERT-Tr. [30]	1.41 M	1.22 M	8.83 G	5.8 ms	84.6%	94.1%
(+RM)	13.96 M	1.30 M	9.42 G	9.6 ms		
ALERT-Tr.	0.04 M	0.0040 M	0.03 G	<b>3.9</b> ms	72.6%	89.2%
(+LMM)	0.57 M	0.0074 M	0.06 G	6.0 ms		
EVA (ours)	0.62 M	0.60 M	4.79 G	14.7 ms	<b>92.9%</b>	<b>96.9%</b>
(+ResNet-14)	2.83 M	0.28 M	1.82 G	<b>1.5</b> ms		

Table 2: Results on N-Cars. EVA (+ResNet-14) is pretrained on N-Cars, and EVA (+ResNet-34) is pretrained on Gen1.

Model	Event Repr.	Acc.
GET [34]	Event count $\rightarrow$ Token	96.9%
MEM [25]	Event count	<b>98.4%</b>
EST [12]	Learned	92.5%
Matrix-LSTM [5]	Learned	95.8%
ALERT-Tr. (+LMM)	Learned	85.6%
EVA (+ResNet-14) (ours)	Learned	91.4%
EVA (+ResNet-14) ( $D_{\text{head}} = 16$ ) (ours)	Learned	91.6%
EVA-L (+ResNet-34) ( $D_{\text{head}} = 16$ ) (ours)	Learned	96.3%

## 4.2 Object detection

**Dataset.** We evaluate our work on the challenging automotive detection dataset Gen1 [8], which includes 39 hours of data recorded by the Gen1 event camera with a sensor resolution of  $304 \times 240$ .

**Implementations.** We present EVA of 3 RWKV-6 blocks, one MVHS layer with  $D = 128$ . We also present a larger model with  $D = 192$  (referred to as EVA-L). We use patch size  $P = 16$  and  $D_{\text{head}} = 16$  to keep the spatial resolution. Like N-Cars, we use only half of the output channels of the MVHS layer. For the downstream detection task, we use RVT-B [14], a state-of-the-art (SOTA) Transformer-based synchronous backbone and YOLOX detection head [11] on the learned representations. The model implementation details are given in Appendix C.

We pretrain the encoder with sequence length  $T = 2048$ , chunk length  $T_{\text{chunk}} = 16$  and batch size 64 on 4 RTX 3090 GPUs. The learning rate is initially 0.001 with a decay of  $\gamma = 0.9$  every epoch. We use EC and TS for MRP objectives and EC for NRP. The details are given in Appendix D.

**Results.** Existing A2S work did not report results on object detection tasks. Therefore, here we compare our method with (i) synchronous tensor-based methods and (ii) end-to-end asynchronous methods. The results are shown in Table 3. End-to-end asynchronous methods could predict results in an event-by-event manner, but have a lower accuracy compared with synchronous methods. Our best model achieves a remarkable 47.7 mAP, and improves the SOTA synchronous method [14] with a lower number of channels of the event representation (6 for our EVA compared with 20 for RVT-B). This is the first result by A2S methods on event-based detection tasks, demonstrating the effectiveness of our method.

Table 3: Object detection performance on Gen1. Async. refers to end-to-end asynchronous methods. Sync. refers to synchronous tensor-based methods.

Model	Type	mAP (%)
NVS-S [28]	Async.	8.6
AEGNN [40]	Async.	14.5
Asynet [31]	Async.	16.3
FARSE-CNN [39]	Async.	30.0
ASTMNet [27]	Sync.	46.7
HMNet-L3 [15]	Sync.	47.1
RVT-B [14]	Sync.	47.2
GET [34]	Sync.	<b>47.9</b>
EVA (+RVT-B) (ours) ( $D = 128, D_{\text{head}} = 16$ )	A2S	47.5
EVA-L (+RVT-B) (ours) ( $D = 192, D_{\text{head}} = 16$ )	A2S	<u>47.7</u>



### 4.3 Ablation study

We ablate architecture components, including MVHS and temporal embedding, and SSL objectives to check their contribution. The effect of patch size is also investigated. The ablation is conducted on DVS128-Gesture, following the implementations in the object recognition experiment. We report the loss items of SSL during pretraining and downstream task accuracy, including FVA and SA, on the validation set.

**Ablation of MVHS and temporal embedding.** To remove MVHS, we keep the model width  $D = 128$  unchanged and let  $D_{\text{head}} = 1$  in the output layer. The output becomes a vector with shape  $128 \times 1 \times 1$ . The results are shown in Tab 4. Models without MVHS or temporal embedding show a decrease in both FVA and SA, and also have a larger pretraining loss.

Table 4: Ablation of MVHS and temporal embedding.

MVHS	Tempo. Emb.	MRP Loss		NRP Loss	Acc.	
		EC ( $\downarrow$ )	TS ( $\downarrow$ )	EC( $\downarrow$ )	FVA ( $\uparrow$ )	SA ( $\uparrow$ )
✓	✓	<b>0.366</b>	<b><math>3.16 \times 10^{-3}</math></b>	<b>0.767</b>	<b>98.1%</b>	<b>94.7%</b>
✓		1.870	$4.56 \times 10^{-2}$	0.933	87.8%	81.1%
	✓	0.826	$9.34 \times 10^{-3}$	0.834	97.4%	94.1%

**Ablation of SSL objective.** We remove the SSL objectives to show their contribution to the learned representation. Results in Table 5 show that removing the items of SSL decreases the performance of the model. One interesting thing is that learning only one representation, such as EC, will not contribute to a smaller pretraining loss compared with learning all the representations. This indicates that learning one single representation could benefit from learning other representations.

Table 5: Ablation of objectives of MRP and NRP.

MRP		NRP	MRP Loss		NRP Loss	Acc.	
EC	TS		EC ( $\downarrow$ )	TS ( $\downarrow$ )	EC( $\downarrow$ )	FVA ( $\uparrow$ )	SA ( $\uparrow$ )
✓	✓	✓	<b>0.366</b>	<b><math>3.16 \times 10^{-3}</math></b>	<b>0.767</b>	<b>98.1%</b>	<b>94.7%</b>
✓	✓		0.639	$1.03 \times 10^{-2}$	–	96.8%	94.4%
✓			0.701	–	–	96.8%	91.7%
	✓		–	$1.40 \times 10^{-2}$	–	97.4%	91.6%
		✓	–	–	0.825	96.8%	91.8%

**Effect of patch size.** We change the patch size from 16 to 128 (w/o patching in this case), while keeping the width of the encoder  $D = 128$  and the sequence length  $T = 8192$  for different patch sizes. The results in Table 6 show that a small patch size gives better accuracy. Patch size  $P = 128$  gives a smaller pretraining loss because the spatial sparsity makes a larger patch have more blank areas.

Table 6: Effect of the choice of patch size

Patch Size	MRP Loss		NRP Loss	Acc.	
	EC ( $\downarrow$ )	TS ( $\downarrow$ )	EC( $\downarrow$ )	FVA ( $\uparrow$ )	SA ( $\uparrow$ )
16	0.366	<b><math>3.16 \times 10^{-3}</math></b>	0.767	<b>98.1%</b>	<b>94.7%</b>
32	0.454	$1.67 \times 10^{-2}$	0.393	94.8%	91.0%
64	0.233	$1.45 \times 10^{-2}$	0.149	95.5%	89.6%
128 (w/o patching)	<b>0.108</b>	$9.88 \times 10^{-3}$	<b>0.073</b>	97.4%	89.3%

## 5 Conclusion

We introduce a new A2S framework, EVA, for real-time event camera data processing. The framework consists of an efficient and expressive linear attention-based asynchronous encoder for event-by-event representation updating and an SSL method to learn a generalizable event representation. EVA outperforms the previous A2S work [30], achieves remarkable performance on event-based detection tasks, and learns generalizable representations applicable in diverse downstream tasks, demonstrating its potential towards real-time event-based applications.

The limitations of our work are discussed in Appendix G.

## References

- [1] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7243–7252, 2017.
- [2] H. Bao, L. Dong, S. Piao, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [3] S. Barchid, J. Mennesson, and C. Djéraba. Exploring joint embedding architectures and data augmentations for self-supervised representation learning in event-based vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3912, 2023.
- [4] M. Cannici, M. Ciccone, A. Romanoni, and M. Matteucci. Asynchronous convolutional networks for object detection in neuromorphic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [5] M. Cannici, M. Ciccone, A. Romanoni, and M. Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 136–152. Springer, 2020.
- [6] S. Chen, S. Wong, L. Chen, and Y. Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.
- [7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmlR, 2020.
- [8] P. De Tournemire, D. Nitti, E. Perot, D. Migliore, and A. Sironi. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*, 2020.
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [10] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.
- [11] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.
- [12] D. Gehrig, A. Loquercio, K. G. Derpanis, and D. Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019.
- [13] D. Gehrig and D. Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024.

- [14] M. Gehrig and D. Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13884–13893, 2023.
- [15] R. Hamaguchi, Y. Furukawa, M. Onishi, and K. Sakurada. Hierarchical neural memory network for low latency event processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22867–22876, 2023.
- [16] B. He, Z. Wang, Y. Zhou, J. Chen, C. D. Singh, H. Li, Y. Gao, S. Shen, K. Wang, Y. Cao, et al. Microsaccade-inspired event camera for robotics. *Science Robotics*, 9(90):eadj8124, 2024.
- [17] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.
- [18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [19] W. He, J. Zhu, Y. Feng, F. Liang, K. You, H. Chai, Z. Sui, H. Hao, G. Li, J. Zhao, et al. Neuromorphic-enabled video-activated cell sorting. *Nature Communications*, 15(1):10792, 2024.
- [20] S. U. Innocenti, F. Becattini, F. Pernici, and A. Del Bimbo. Temporal binary representation for event-based action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10426–10432. IEEE, 2021.
- [21] U. Kamal, S. Dash, and S. Mukhopadhyay. Associative memory augmented asynchronous spatiotemporal representation learning for event-based perception. In *The Eleventh International Conference on Learning Representations*, 2023.
- [22] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.
- [23] A. Kendall, Y. Gal, and R. Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.
- [24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [25] S. Klenk, D. Bonello, L. Koestler, N. Araslanov, and D. Cremers. Masked event modeling: Self-supervised pretraining for event cameras. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2378–2388, 2024.
- [26] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. B. Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016.
- [27] J. Li, J. Li, L. Zhu, X. Xiang, T. Huang, and Y. Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing*, 31:2975–2987, 2022.
- [28] Y. Li, H. Zhou, B. Yang, Y. Zhang, Z. Cui, H. Bao, and G. Zhang. Graph-based asynchronous event processing for rapid object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 934–943, 2021.
- [29] E. Martin and C. Cundy. Parallelizing linear recurrent neural nets over sequence length. In *International Conference on Learning Representations*, 2018.
- [30] C. Martin-Turrero, M. Bouvier, M. Breitenstein, P. Zanuttigh, and V. Parret. Alert-transformer: Bridging asynchronous and synchronous machine learning for real-time event-based spatio-temporal data. *arXiv preprint arXiv:2402.01393*, 2024.

- [31] N. Messikommer, D. Gehrig, A. Loquercio, and D. Scaramuzza. Event-based asynchronous sparse convolutional networks. In *European Conference on Computer Vision*, pages 415–431. Springer, 2020.
- [32] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. Chung, M. Grella, et al. Rwkv: Reinventing rnnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.
- [33] B. Peng, D. Goldstein, Q. Anthony, A. Albalak, E. Alcaide, S. Biderman, E. Cheah, T. Ferdinan, H. Hou, P. Kazienko, et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 3, 2024.
- [34] Y. Peng, Y. Zhang, Z. Xiong, X. Sun, and F. Wu. Get: Group event transformer for event-based vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6038–6048, 2023.
- [35] O. Press, N. A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.
- [36] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.
- [37] Q. Qu, X. Chen, Y. Y. Chung, and Y. Shen. Evrepsl: Event-stream representation via self-supervised learning for event-based vision. *IEEE Transactions on Image Processing*, 2024.
- [38] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training.(2018), 2018.
- [39] R. Santambrogio, M. Cannici, and M. Matteucci. Farse-cnn: Fully asynchronous, recurrent and sparse event-based cnn. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.
- [40] S. Schaefer, D. Gehrig, and D. Scaramuzza. Aegnn: Asynchronous event-based graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12371–12381, 2022.
- [41] Y. Sekikawa, K. Hara, and H. Saito. Eventnet: Asynchronous recursive event processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3887–3896, 2019.
- [42] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1731–1740, 2018.
- [43] T. Soydan, N. Zubic, N. Messikommer, S. Mishra, and D. Scaramuzza. S7: Selective and simplified state space layers for sequence modeling. *ArXiv*, abs/2410.03464, 2024.
- [44] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- [45] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, and F. Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.
- [46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [47] S. Yang, B. Wang, Y. Shen, R. Panda, and Y. Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.
- [48] Y. Yang, L. Pan, and L. Liu. Event camera data pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10699–10709, 2023.

- [49] Y. Yang, L. Pan, and L. Liu. Event camera data dense pre-training. In *European Conference on Computer Vision*, pages 292–310. Springer, 2024.
- [50] N. Zubić, D. Gehrig, M. Gehrig, and D. Scaramuzza. From chaos comes order: Ordering event representations for object recognition and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12846–12856, October 2023.
- [51] N. Zubic, M. Gehrig, and D. Scaramuzza. State space models for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5819–5828, June 2024.
- [52] N. Zubic and D. Scaramuzza. Gg-ssms: Graph-generating state space models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

## A Architecture details of the asynchronous encoder

**RWKV-6 details.** As described in [33], one RWKV-6 block is composed of token mixing (TM) and channel mixing (CM). For input sequence  $\{\mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^T$  of length  $T$ , TM first conduct a data-dependent linear interpolation (ddlerp) between  $\mathbf{x}_i$  and  $\mathbf{x}_{i-1}$  by

$$\begin{aligned} \text{lora}_{\{r,k,v,g\}}(\mathbf{x}_i) &= \boldsymbol{\lambda}_{\{r,k,v,g\}} + \tanh(\mathbf{x}_i \mathbf{A}_{\{r,k,v,g\}}) \mathbf{B}_{\{r,k,v,g\}}, \\ \text{ddlerp}_{\{r,k,v,g\}}(\mathbf{x}_i, \mathbf{x}_{i-1}) &= \mathbf{x}_i + (\mathbf{x}_{i-1} - \mathbf{x}_i) \odot \text{lora}_{\{r,k,v,g\}}(\mathbf{x}_i + (\mathbf{x}_{i-1} - \mathbf{x}_i) \odot \boldsymbol{\mu}), \end{aligned} \quad (8)$$

where  $\boldsymbol{\mu}, \boldsymbol{\lambda}_{\{r,k,v,g\}} \in \mathbb{R}^D$  are learnable parameters and  $\mathbf{A}_{\{r,k,v,g\}} \in \mathbb{R}^{D \times D_{\text{lora}}}$ ,  $\mathbf{B}_{\{r,k,v,g\}} \in \mathbb{R}^{D_{\text{lora}} \times D}$  are trainable weight matrices. The vectors in Equation 4 are calculated as

$$\begin{aligned} \mathbf{r}_t &= \text{ddlerp}_r(\mathbf{x}_i, \mathbf{x}_{i-1}) \mathbf{W}_r, \\ \mathbf{k}_t &= \text{ddlerp}_k(\mathbf{x}_i, \mathbf{x}_{i-1}) \mathbf{W}_k, \\ \mathbf{v}_t &= \text{ddlerp}_v(\mathbf{x}_i, \mathbf{x}_{i-1}) \mathbf{W}_v, \\ \mathbf{g}_t &= \text{ddlerp}_g(\mathbf{x}_i, \mathbf{x}_{i-1}) \mathbf{W}_g, \\ \mathbf{d}_t &= \text{lora}_d(\text{ddlerp}_d(\mathbf{x}_i, \mathbf{x}_{i-1})), \\ \mathbf{w}_t &= \exp(-\exp(\mathbf{d}_t)), \end{aligned} \quad (9)$$

where  $\mathbf{W}_{\{r,k,v,g\}} \in \mathbb{R}^{D \times D}$ ,  $\mathbf{A}_w \in \mathbb{R}^{D \times D_w}$ ,  $\mathbf{B}_w \in \mathbb{R}^{D_w \times D}$  are trainable weight matrices. By Equation 4 we have  $\mathbf{y}_i \in \mathbb{R}^{D \times 1}$ . The output of TM is given as

$$\mathbf{o}_t = \text{concat}(\text{SiLU}(\mathbf{g}_t) \odot \text{LayerNorm}(\mathbf{y}_t)) \mathbf{W}_o \in \mathbb{R}^{D \times 1}. \quad (10)$$

In CM, for inputs  $\{\mathbf{x}'_i \in \mathbb{R}^D\}_{i=1}^T$ , a linear interpolation (lerp) is first conducted as

$$\text{lerp}_{\{r',k',v'\}}(\mathbf{x}'_i, \mathbf{x}'_{i-1}) = \mathbf{x}'_i + (\mathbf{x}'_{i-1} - \mathbf{x}'_i) \boldsymbol{\mu}_{\{r',k',v'\}}, \quad (11)$$

where  $\boldsymbol{\mu}_{\{r',k',v'\}} \in \mathbb{R}^D$  is trainable parameter. CM is conducted as

$$\begin{aligned} \mathbf{r}'_t &= \text{lerp}_{r'}(\mathbf{x}'_i, \mathbf{x}'_{i-1}) \mathbf{W}_{r'} \in \mathbb{R}^D, \\ \mathbf{k}'_t &= \text{lerp}_{k'}(\mathbf{x}'_i, \mathbf{x}'_{i-1}) \mathbf{W}_{k'} \in \mathbb{R}^{D_{\text{fm}}}, \\ \mathbf{v}'_t &= \text{ReLU}(\mathbf{k}'_t)^2 \mathbf{W}_{v'} \in \mathbb{R}^D, \\ \mathbf{o}'_t &= \sigma(\mathbf{r}'_t) \odot \mathbf{v}'_t \in \mathbb{R}^D, \end{aligned} \quad (12)$$

where  $\mathbf{W}_{r'} \in \mathbb{R}^{D \times D}$ ,  $\mathbf{W}_{k'} \in \mathbb{R}^{D \times D_{\text{fm}}}$ ,  $\mathbf{W}_{v'} \in \mathbb{R}^{D_{\text{fm}} \times D}$  are trainable matrices.

**MVHS layer.** The MVHS layer is modified from the TM layer in RWKV-6. For input sequence  $\{\mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^T$  of length  $T$ , MVHS layer first maps them into  $\mathbf{k}_t, \mathbf{v}_t, \mathbf{w}_t$  as Equation 9. Then, given number of heads  $N$  and dimension of head  $D_{\text{head}} = D/N$ , the output  $\mathbf{S}$  of the MVHS layer is given

as

$$\begin{aligned}
\mathbf{w}_i^h &= \mathbf{w}_i [hD_{\text{head}} : (h+1)D_{\text{head}} - 1] \in \mathbb{R}^{D_{\text{head}}}, \\
\mathbf{k}_i^h &= \mathbf{k}_i [hD_{\text{head}} : (h+1)D_{\text{head}} - 1] \in \mathbb{R}^{D_{\text{head}}}, \\
\mathbf{v}_i^h &= \mathbf{v}_i [hD_{\text{head}} : (h+1)D_{\text{head}} - 1] \in \mathbb{R}^{D_{\text{head}}}, \\
\mathbf{S}_i^h &= \sum_{t=1}^i \text{diag} \left( \bigotimes_{j=t+1}^i \mathbf{w}_j^h \right) \mathbf{k}_t^h (\mathbf{v}_t^h)^T \in \mathbb{R}^{D_{\text{head}} \times D_{\text{head}}}, h = 0, 1, \dots, N-1, \\
\mathbf{S}_i &= \text{concat}(\{\mathbf{S}_i^h\}_{h=0}^{N-1}) \in \mathbb{R}^{N \times D_{\text{head}} \times D_{\text{head}}}.
\end{aligned} \tag{13}$$

This is equivalent to directly using the multi-channel MVHS of the TM in RWKV-6 as the output. The hidden state preserves the aggregated information of the input events, which is precisely what we want for event representation. Moreover, the multi-channel matrix shape matches the multi-channel image-like handcrafted event representations and could convey expressive fine-grained spatial information. Therefore, we think using MVHS as output is a key towards expressive event representation.

## B Dataset preprocess

**DVS128-Gesture.** DVS128-Gesture has 11 classes of human actions of 29 different users. The official split of the dataset uses users 0 to 23 as the train split and users 24 to 29 as the test split. We follow this method and use users 21 to 23 in the original train split as a validation set. The pretraining is conducted on the new training set (user 0 to 20) and validated on the val set (user 21 to 23). We report the test accuracy of the best model on the validation set.

For pretraining the model on sequences of length 8192, we first split all the files into patches of patch size  $P = 16$ . Then, we slice each patch into sequences of length 10240 with a stride of 8192 as samples. In each sample, the first 8192 events are the inputs to the model, and the last 2048 events are used to generate the NRP objectives. We save the samples in the data format of  $\Delta t, x, y, p$  and convert them into uint16 to save storage. The difference of timestamps  $\Delta t$  may be clamped to the numeric range of uint16, but this happens seldom and has little influence on the experiments.

**N-Cars.** N-Cars has files of duration 100 ms. Each file has around 4000 events. After splitting the files into patches of size  $16 \times 16$ , the average length of each patch is around 200, which is quite small. Therefore, on this dataset, we split the patches into samples of length 768 with a stride of 512. The first 512 events are inputs, and the last 256 events are used for NRP objectives. Following [12, 5], we use 20% of the training set as the validation split. The samples are saved in the same format as DVS128-Gesture.

**Gen1.** We first filter out all the hot pixels caused by event camera sensor errors in the dataset. For every 10 ms in the file, the events in a pixel are deleted if they activate more than 40 times (we check the normal files and find this value is no more than 10 regularly). This is similar to limiting the maximum value in the event counts in the synchronous method [14]. We use the official train-val-test split of Gen1 and split the files into patches of size  $16 \times 16$ . Since Gen1 is a large dataset with sufficient events for pretraining the model, we do not use all the events for pretraining in one epoch. Instead, we uniformly sample a slice of length 4096 from each patch. The first 2048 events are inputs, and the last 2048 events are left for NRP objectives. For pretraining, we save the patches in the same format as DVS128-Gesture. For training, all the generated representations are multiplied by a factor of 1/8, converted to their absolute value of data type uint8 to be consistent with the numerical range of the original inputs of RVT (0 to 20). We do this because we find that if we use the original float data type of the representation, there will probably be a numerical NAN error caused by the YOLOX head [11]. This could also help save storage.

## C Implementations

Three asynchronous encoders are presented in the experiment. In the experiments on DVS128-Gesture, we present an encoder (EVA) with 3 RWKV-6 blocks and 1 MVHS layer with  $D = 128$ ,  $D_{\text{head}} = 8$ ,  $N = 16$ ,  $D_{\text{ffn}} = 256$ ,  $D_{\text{lora}} = D_w = 16$ . The shape of the MVHS output is  $(16 \times 8 \times 8)$ . To predict the representation of shape  $(* \times 16 \times 16)$ , we use a ResNet head with 1 basic ResNet block of width 64, 1 ConvTranspose2d( $in = 64$ ,  $out = 32$ ,  $ks = 4$ ,  $stride = 2$ ) to upsample the  $8 \times 8$

hidden to  $16 \times 16$  and 1 basic ResNet block of width 32. We use one  $1 \times 1$  convolution layer as output. In the experiment on N-Cars, except for the encoder (EVA) on DVS128-Gesture, we also present an encoder with  $N = 8, D_{\text{head}} = 16$  (referred to as EVA ( $D_{\text{head}} = 16$ ) in the main context) since the spatial resolution of files in N-Cars is already cropped and small in spatial resolution. Using a shrunk hidden state as representation could indeed help reduce downstream tasks inference latency, but will decrease the accuracy (see Table 2). To keep the same total size of the representation, we use only half the channels in the MVHS layer, that is  $N = 4, D_{\text{head}} = 16$ . The shape of the MVHS output is  $4 \times 16 \times 16$ . In this case, since there is no need for upsampling, we use a ResNet with 1 basic ResNet block of width 32 as the head. In the experiment on Gen1, we also present a larger encoder with  $D = 192, D_{\text{head}} = 16, N = 8, D_{\text{ffn}} = 384$  (referred to as EVA-L ( $D = 192, D_{\text{head}} = 16$ ) in the main context) in the RWKV-6 blocks and  $N = 6, D_{\text{head}} = 16$  in the MVHS layer. The same ResNet head as N-Cars is used for pretraining.

For initialization, we use the official initialization method in RWKV-6 for the encoder and Kaiming initialization for the ResNet. The asynchronous encoder, including the RWKV-6 blocks and MVHS layer, is trained with bfloat16 precision as the default value in the original implementations in RWKV-6.

In the recognition task on DVS128-Gesture, we use a 14-layer (ResNet-14) as the classifier by removing the last stage of the original ResNet-18 [18] to reduce inference latency (see Table 7). Interestingly, ResNet-14 has a higher accuracy on the validation set with a smaller parameter size and lower latency. We think this is because a smaller model could help overcome overfitting. Despite this, we find that ResNet-18 could also outperform previous work [30]. We use batch size 32, Adam optimizer with learning rate 0.001, mixup augmentation, and train 200 epochs on a single RTX 3090 GPU with a decay of learning rate  $\gamma = 0.8$  every 20 epochs. In the experiments on N-Cars, we follow

Table 7: Classification results on DVS128-Gesture with ResNet-18

Model	Classifier	# Params	Latency	Val. acc. (SA)	Test acc. (SA)
ALERT-Tr. [30]	RM	13.96 M	9.6 ms	—	84.6%
EVA	ResNet-18	11.28 M	2.0 ms	92.8%	92.0%
EVA	ResNet-14 (used)	2.83 M	1.5 ms	94.1%	92.9%

[5, 12] to use a ResNet-34 pretrained on ImageNet as the classifier. Following the train settings of previous work, we use a batch size of 64, Adam optimizer with a constant learning rate of 0.0001, random flip as data augmentation, and train 100 epochs on an RTX 3090 GPU. In the experiments on Gen1, we use the default training setups of RVT[14] for a fair comparison. We use batch size 8, learning rate  $2e-4$ , and train 400,000 steps on a single A800 GPU. The training takes about 2 days.

## D Pretraining objectives

We use event count (EC) and time surface (TS) as the pretraining objectives. Consider events  $\mathcal{E} = \{e_i = (t_i, x_i, y_i, p_i)\}_{i=1}^T$  and patch size  $P$ , EC of a given time window  $[t_s, t_e]$  is constructed as a 2 channel image of shape  $2 \times P \times P$  by

$$\text{EC}(p, x, y) = \sum_i 1_{x_i=x, y_i=y, p_i=p, t_i \in [t_s, t_e]} \in \mathbb{R}^{2 \times P \times P}. \quad (14)$$

and TS is calculated as

$$\text{TS}(p, x, y) = \max_i \exp\left(\frac{t_i - t_T}{\tau}\right) \times 1_{x_i=x, y_i=y, p_i=p} \in \mathbb{R}^{2 \times P \times P}, \quad (15)$$

where  $\tau > 0$  is a time constant.

The objectives in the experiments are given in Table 8. On Gen1, we use multiple ECs and larger TS  $\tau$  because we find it could help converge at the initial period of pretraining to decrease the time required to pretrain on the large dataset.

For a better understanding of the effects of the SSL objectives, here we give a study on DVS128-Gesture. The results are given in Table 9. We change the SSL objectives and find that the validation set accuracy (SA) is not heavily affected. The optimal configuration of the hyperparameter has not been investigated yet, and we hope to leave it for future study.

Table 8: Objectives for the SSL. For EC in MRP

Dataset	MRP		NRP
	EC time window (ms)	TS $\tau$ (ms)	EC time window (ms)
DVS128-Gesture	100	100	20
N-Cars	100	100	20
Gen1	50, 25, 10, 5	200	10

Table 9: Effect of SSL objectives on DVS128-Gesture

MRP		NRP	MRP Loss		NRP Loss	Acc.	
EC (ms)	TS (ms)	EC (ms)	EC	TS	EC	FVA	SA
100	100	20	0.366	$3.16 \times 10^{-3}$	0.767	98.1%	94.7%
100 $\rightarrow$ 50	100	20	0.200	$2.33 \times 10^{-3}$	0.764	95.5%	94.3%
100 $\rightarrow$ 200	100	20	0.807	$3.60 \times 10^{-3}$	0.767	95.5%	94.3%
100	100 $\rightarrow$ 50	20	0.398	$2.39 \times 10^{-3}$	0.766	98.1%	94.1%
100	100 $\rightarrow$ 200	20	0.374	$3.78 \times 10^{-3}$	0.780	96.2%	94.4%
100	100	20 $\rightarrow$ 10	0.389	$3.29 \times 10^{-3}$	0.364	97.4%	94.2%
100	100	20 $\rightarrow$ 40	0.353	$2.86 \times 10^{-3}$	1.698	98.7%	94.3%

## E Visualization

We visualize the handcrafted representations for SSL pretraining, the predicted results of the pretraining target, and the learned representations in Figure 5. The results are generated on DVS128-Gesture. The comparison between Figure 5 (a) and (b) demonstrates that during SSL, our model could consistently predict multiple handcrafted representations (EC, TS) of the events with high similarity. Additionally, it shows favorable visualization results in predicting the EC for future time periods. This indicates that our model is capable of predicting short-term motion patterns rather than merely memorizing the history.

## F Details of the MVHS and patch-wise encoding to reduce model size

Assume that we keep the size of the learned representation unchanged. For original vector-value output, the output size is  $D$ , and the model parameter size is around  $\mathcal{O}(D^2)$  (since the majority of the parameters come from the matrix). For MVHS output, the output size is  $N \times (D'/N) \times (D'/N)$  with model size  $\mathcal{O}(D'^2)$ . Let  $D = N \times (D'/N) \times (D'/N)$ , we have  $\mathcal{O}(D^2)/\mathcal{O}(D'^2) = D/N = D_{\text{head}}$ . In our EVA, this factor is 8 or 16.

For patch-wise encoding, assume that we learn a representation that is proportional to the size of the objective representations. Consider an event camera with resolution  $(H, W)$  (assume  $H = W$  for simplicity) and patch size  $P$ , the number of patches  $\# \text{ patches}$  is  $H^2/P^2$ . Given the shape of the learned representation as  $(N \times D \times D)$ , if we use it to predict the objective representation of full size, which is  $(* \times H \times H)$ , then we have  $D = H$  and the model size is around  $\mathcal{O}(H^2)$ . However, if we use the learned representation to predict a the representations of a patch, which have the shape of  $(* \times P \times P)$ , we have  $D = P$ , and the model size is around  $\mathcal{O}(P^2)$ . Therefore, the model size decreases  $\mathcal{O}(H^2)/\mathcal{O}(P^2) = \# \text{ patches}$  times. In the experiment on DVS128-Gesture, we have  $H = W = 128$  and  $P = 16$ . This factor is 64. On Gen1 with  $H = 240, W = 304$  and  $P = 16$ , we have  $\# \text{ patches} = 285$ .

## G Limitations

Our work is limited in two aspects. First, while our EVA framework is evaluated on publicly available datasets recorded in real-world settings, especially in automotive driving scenarios, a full exploration of its utility in practical applications requires implementing the models on device-side, such as on FPGAs, and testing the whole framework in actual deployment scenarios. These aspects are not



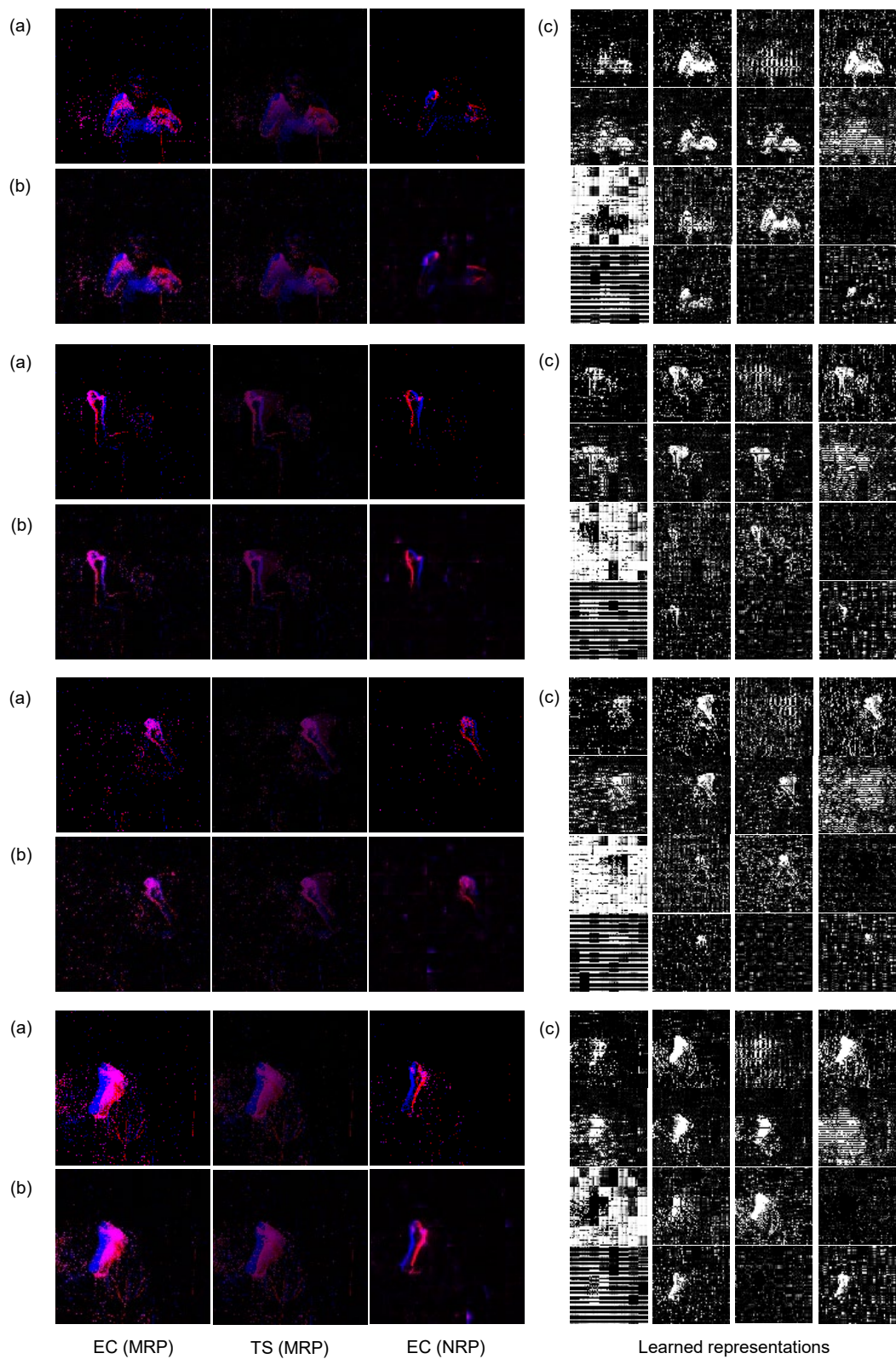


Figure 5: Visualization on DVS128-Gesture. (a) Ground truth of the handcrafted representations. (b) Predicted results of the handcrafted representations. (c) The learned representations.

involved in the current study. Secondly, the models trained are relatively small in parameter size compared to pretrained large language models (LLMs) with billions of parameters. Consequently, the potential of self-supervised pretraining could not be fully exploited in our study. The parameter size of the asynchronous encoder is constrained by the requirement for real-time processing of high-temporal-resolution events. We aim to address this limitation in the future by decreasing the per-event computational complexity, thereby enabling the design of larger pretrained models capable of asynchronously processing events in real time.