Computer Vision and Image Understanding 264 (2026) 104619

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

EventSleep2: Sleep activity recognition on complete night sleep recordings with an event camera ☆

Nerea Gallego [a,*,1], Carlos Plou [a,1], Miguel Marcos [a], Pablo Urcola [b], Luis Montesano [a,b], Eduardo Montijano [a], Ruben Martinez-Cantin [a], Ana C. Murillo [a]

[a] DIIS - Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Spain
[b] Bitbrain Technologies, Zaragoza, Spain

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Event cameras
Sleep activity recognition
Low-light conditions
Bayesian deep learning

ABSTRACT

Sleep is fundamental to health, and society is more and more aware of the impact and relevance of sleep disorders. Traditional diagnostic methods, like polysomnography, are intrusive and resource-intensive. Instead, research is focusing on developing novel, less intrusive or portable methods that combine intelligent sensors with activity recognition for diagnosis support and scoring. Event cameras offer a promising alternative for automated, in-home sleep activity recognition due to their excellent low-light performance and low power consumption. This work introduces **EventSleep2-data**, a significant extension to the EventSleep dataset, featuring 10 complete night recordings (around 7 h each) of volunteers sleeping in their homes. Unlike the original short and controlled recordings, this new dataset captures natural, full-night sleep sessions under realistic conditions. This new data incorporates challenging real-world scene variations, an efficient movement-triggered sparse data recording pipeline, and synchronized 2-channel EEG data for a subset of recordings. We also present **EventSleep2-net**, a novel event-based sleep activity recognition approach with a dual-head architecture to simultaneously analyze motion classes and static poses. The model is specifically designed to handle the motion-triggered, sparse nature of complete night recordings. Unlike the original EventSleep architecture, EventSleep2-net can predict both movement and static poses even during long periods with no events. We demonstrate state-of-the-art performance on both EventSleep1-data, the original dataset, and EventSleep2-data, with comprehensive ablation studies validating our design decisions. Together, EventSleep2-data and EventSleep2-net overcome the limitations of the previous setup and enable continuous, full-night analysis for real-world sleep monitoring, significantly advancing the potential of event-based vision for sleep disorder studies. Code and data are publicly available on the webpage: https://sites.google.com/unizar.es/eventsleep.

1. Introduction

Sleep accounts for about a third of human lives and plays a key role in maintaining our physical and mental health. Cases of sleep disorder are quite pervasive, with up to 50% of the adult population claiming to suffer from some problem (Vanderlinden et al., 2020) and approximately 34% among children and teenagers (Cai et al., 2024).

Given the broad impact on the population and increasing awareness of related disorders, sleep is increasingly being studied through diverse sensor technologies.

There are multiple phone applications and commodity gadgets that offer sleep analysis. These applications use sensors such as microphones, IMUs, and pulse oxymeters, which are common in phones and smart watches, to compute sleep phases and events. This type of applications provide results which are far from the accuracy that can be obtained in medical studies using more complex setups, in particular those including EEG data (Lee et al., 2023). Our work scope is not commodity gadgets, but systems that can bring medical studies to the patients' homes.

One crucial goal in this direction is to ease the assessment and monitoring of sleep conditions in more natural and comfortable settings (Moyen et al., 2024) compared to traditional, lab-based polysomnography (PSG). This is particularly beneficial for long-term monitoring of chronic pathologies that often produce sleep disturbances. From wearables, such as the headband from López-Larraz et al. (2023), to

☆ This article is part of a Special issue entitled: 'ACVR 2024' published in Computer Vision and Image Understanding.

* Corresponding author.
E-mail addresses: ngallego@unizar.es (N. Gallego), c.plou@unizar.es (C. Plou).

[1] Nerea Gallego and Carlos Plou contributed equally in this work.

https://doi.org/10.1016/j.cviu.2025.104619
Received 16 May 2025; Received in revised form 1 November 2025; Accepted 17 December 2025
Available online 2 January 2026

Fig. 1. Overview of EventSleep2 – The new EventSleep2-data contains 10 complete night recordings or people sleeping in their homes, and 10 activity labels including motion and pose classes. The new EventSleep2-net, with specific heads for Motion and Static classes, updates EventSleep1-net approach to enable processing the complete night recordings.

contactless systems, such as camera-based approaches from Akbarian et al. (2021), Plou et al. (2024), these technologies are providing researchers with new insights into sleep activity.

While sleeping, we perform all kinds of movements such as kicks, rolls, etc. Certain abnormal movements are symptoms of sleep disorders such as restless legs syndrome, insomnia, or sleep apnea, and can help identify them in a monitored patient. However, the traditional diagnosis of sleep pathologies is based on polysomnography, which relies on experts to manually determine and classify the presence of disorders (Hassan and Subasi, 2017; Huang et al., 2014). To alleviate these costly manual processes, we study the suitability of event-based sensors for the automated detection and recognition of subjects' movements during their sleep. Event-based cameras can complement the information that other sensors provide (e.g., EEG, ECG, SpO2...), whose performance can be affected by movement during sleep. Moreover, event cameras excel where traditional RGB cameras would pose a hindrance. Specifically, (1) they work well in pitch-dark environments, (2) provide a very dense temporal resolution, (3) better maintain the privacy of the subject compared to RGB or infrared cameras, and (4) require less power, as they are only triggered by the sparse, intermittent movements of the subject.

In order to facilitate the development of applications for real scenarios, this work extends our pioneering EventSleep dataset (Plou et al., 2024) (hereafter referred to as EventSleep1-data and EventSleep1-net for clarity). Unlike the previous version, which was conceived as a proof of concept based on short, controlled recordings, the present work focuses on realistic long-term monitoring under more natural conditions. EventSleep1-data featured short laboratory sessions where volunteers enacted isolated sleep movements, limiting the realism and continuity of the recordings. In contrast, EventSleep2-data extends it with 10 complete night sleep recordings of a few volunteers while sleeping at night at their homes, bringing not only much more data, but also the following additions.

First, it provides challenging scene variations that better approximate real settings, with data from people sleeping during complete nights. Consequently, it has been collected with a more efficient setup. In order to enable long-duration recordings of up to 8 h while keeping storage and processing requirements manageable, the data collection pipeline has been updated. Instead of raw data collection, the system now filters out background noise and stores only the segments where subject movement occurs. This motion-triggered acquisition preserves all relevant motion information while discarding prolonged inactive periods, allowing realistic, full-night recordings that were not feasible with the original setup.

Besides, EventSleep2-data includes synchronized EEG measurements from a 2-channel neuroheadband for a set of the recorded nights, to encourage further research on the application of multimodal computer vision for sleep disorder study.

Moreover, building on the described new recording strategy, EventSleep2 also introduces a novel event-based recognition approach (EventSleep2-net), specifically designed to handle the challenges of complete night real data recordings. Unlike EventSleep1-net, which was trained on continuous recordings where static poses were treated as an additional activity class, EventSleep2-net is specifically adapted to the new motion-triggered recording procedure, in which only movement clips contain events while static clips remain empty. This adaptation enables the model to handle full-night recordings, generating predictions throughout the night, including during empty static segments where it propagates the last detected pose. EventSleep2-net incorporates a dual-head architecture that specializes each output for *motion* classes, i.e., subject movements, and *static* classes, i.e., subject poses. The motion head is responsible for classifying the six action classes defined in the original dataset, while the static head determines whether to update the subject's static pose. This more natural division enables the system to implicitly distinguish between action and static pose labels, maintaining a consistent estimation of the subject's position throughout the night even during inactivity segments where no event data is recorded.

We evaluate our method on both EventSleep1-data and EventSleep2-data, achieving state-of-the-art results on the complete night sleep sequences. A detailed ablation study further validates several design choices made in our new training and inference pipeline, confirming their contribution to the overall performance of the presented method.

Overall, the key contributions in this work are:

- EventSleep2-data, an extension of the EventSleep1-data dataset that includes 10 sleep recordings of around 7 h each, with a set of synchronized EEG data taken from a neuroheadband. It replaces the controlled short recordings of the previous dataset with natural, long-term recordings enabled by a new motion-triggered acquisition pipeline. This provides unique new data for the sleep research community, in particular for the promising event-camera-based application in this field.
- A discussion of the EEG-event data correlation, emphasizing the correspondence between detected movement and peaks in the EEG signal. These examples provide a stepping stone for future research bridging both modalities.
- EventSleep2-net, a new sleep activity recognition approach based on event camera data. It is built on the EventSleep1-net, but has been redesigned to adapt to the new motion-triggered recordings, which contain long static periods without events. EventSleep2-net advantages are demonstrated with superior performance than previous work on both EventSleep1-data and EventSleep2-data.

2. Related work

2.1. Sleep activity recognition

Sleep activity recognition has been studied with different sensory inputs. Most of the related literature relies on wearable sensors (Yadav

Table 1
`EventSleep2-data` **compared to related event-based activity recognition datasets**, including activity type, modality, resolution, number of classes and subjects, illumination/occlusion conditions, and recording duration.

Dataset	Activity	Modality	Resolution	Classes	Subj.	Dark	Occ.	Clips	Time/Clip	Total time
n-HAR, (Pradhan et al., 2019)	General	Events	304 × 240	5	30	No	No	3091	N/A	N/A
DailyAction, (Liu et al., 2021)	General	Events	128 × 128	12	15	No	No	1440	N/A	N/A
DVS128 Gesture, (Amir et al., 2017)	General	Events+RGB	128 × 128	11	29	No	No	1342	6s	8,052s
THU$^{\text{E-ACT}}$-50, (Gao et al., 2023)	General	Events+RGB	1280 × 800	50	105	No	No	10,500	3.5s	36,750s
THU$^{\text{E-ACT}}$-50-CHL, (Gao et al., 2023)	General	Events+RGB	346 × 260	50	18	*	No	2330	3.5s	8,155s
THU$^{\text{MV-EACT}}$-50, (Gao et al., 2024)	General	Events Multiview	1280 × 800	50	105	No	No	31,500	–	–
DailyDVS-200, (Wang et al., 2024b)	General	Events+RGB	320 × 240	200	47	No	No	22,046	1-20s	–
SeAct, (Zhou et al., 2024)	General	Events+Language	346 × 260	58	–	No	No	–	N/A	N/A
Bullying10K, (Dong et al., 2023)	General	Events	346 × 260	10	25	No	No	10,000	2-20s	–
HARDVS, (Wang et al., 2024a)	General	Events	346 × 260	300	–	*	*	100,000	5s	–
PAF, (Miao et al., 2019)	Office	Events	346 × 260	10	15	No	No	450	5s	2,250s
ASL-DVS, (Bi et al., 2020)	Sign Language	Events	240 × 180	24	5	No	No	100,800	0.1s	10,800s
SL-Animals-DVS, (Vasudevan et al., 2021)	Sign Language	Events	128 × 128	19	59	No	No	1121	4.26s	4,775s
`EventSleep1-data` (Plou et al., 2024)	**Sleep (Simulated)**	**Events+IR**	640 × 480	10	14	**Yes**	**Yes**	1016	5.07s	5,151s
`EventSleep2-data`	**Sleep (Real)**	**Events+EEG**	640 × 480	10	3	**Yes**	**Yes**	10^{+}	7h	70h

*THU$^{\text{E-ACT}}$-50-CHL includes some poorly illuminated clips, but not in low light or dark scenarios as `EventSleep`.
HARDVS includes some multi-illumination and occlusion examples.
Complete and continuous night recordings, while the others contain short clips.

et al., 2021; Zhang et al., 2019) or environmental sensors such as sound (Kay et al., 2012), radio (Liu et al., 2019; Piriyajitakonkij et al., 2021), or light sensors. The majority of datasets that use wearable sensors are based on polysomnograms, which are the standard for sleep quality measurements. However, the process of gathering physiological signals during a polysomnogram, such as an electroencephalogram (EEG), electrocardiogram (ECG), or electromyogram (EMG), requires subjects to be monitored by a fully equipped unit with technicians. There are several public datasets using this kind of setup (Kemp et al., 2000; Terzano et al., 2001; Khalighi et al., 2016), and some works explore bridging other modalities, like sound, with wearable devices like ear-plugged microphones (Han et al., 2024). Although prior work regarding external sensors is limited, despite their interest for unimodal or multimodal activity recognition (Yadav et al., 2021; Sathyanarayana et al., 2018) when complemented with other wearable sensors. When it comes to non-invasive sensors, video data is the principal data modality used. Sleep monitoring, in particular among other image analysis problems, presents the challenge of low illumination (Nakajima et al., 2000). The common approach to overcome this is the use of near-infrared (Mohammadi et al., 2020; Akbarian et al., 2021; Carter et al., 2024) and depth cameras (Carmona et al., 2023). Initially, Mohammadi et al. (2020) employed a single infrared (IR) camera to capture 11 sleep behaviors in 12 participants concurrently with PSG. Additionally, the BlanketSet dataset comprises RGB-IR-D recordings of 8 sleep actions by 14 subjects in a hospital bed (Carmona et al., 2023). Both datasets involve non-clinical healthy participants to enable data sharing and robust labeling, as in our case. Nevertheless, some studies include subjects suffering from sleep conditions, like Akbarian et al. (2021), where they perform sleep apnea classification over IR recordings. Other works propose alternative devices to cameras, such as pressure sensors on the bed, which provide a *heat map* of the subject's body (Matar et al., 2019), radar sensors above the bed that detect position changes (Piriyajitakonkij et al., 2021), or radio emitters and receivers to the sides of the bed that detect signal changes when the subject show abnormal breathing patterns like snoring or coughing (Liu et al., 2019). Differently from all these works, our dataset primary sensor is an event camera.

2.2. Event-based methods for activity recognition

Event cameras record visual information sparsely and asynchronously, generating an event whenever a change in pixel intensity occurs. Each event encodes its spatial location, timestamp, and polarity, resulting in an efficient stream of motion-dependent data rather than continuous frames. Event-based approaches for activity recognition often rely on deep learning, and can be mostly divided into two categories depending on how they process events. On the one hand, some specific architecture designs, like PointNet-like Networks (Wang et al., 2019; Sun et al., 2025), Graph Neural Networks (Bi et al., 2020; Deng et al., 2021), or Spiking Neural Networks (Shrestha and Orchard, 2018; Parameshwara et al., 2021; Vicente-Sola et al., 2025), benefit from the sparsity of event data. PointNet-like networks work with point clouds, which are common representations for sparse data. Similarly, Graph Neural Networks represent events as graph nodes, and neighbor events (in space and time) are connected by edges. Finally, Spiking Neural Networks work through special artificial neurons that mimic the behavior of natural neurons, firing sparse impulses or 'spikes', when their potential reaches certain thresholds, and encode information in the timing of these spikes. On the other hand, some methods preprocess events into dense event representations, or frame representations. Examples of such preprocessing include time surfaces (Lagorce et al., 2016), surfaces of Active Events (Mueggler et al., 2015), binary frame representations (Ghosh et al., 2019; Innocenti et al., 2021), voxel-based representations (Liu et al., 2022), histograms (Sabater et al., 2022), queuing mechanisms (Baldwin et al., 2022) or learned models (Cannici et al., 2020). Among the most common, dense frame representations are usually processed with deep learning models such as CNNs (Amir et al., 2017; Cannici et al., 2020; Innocenti et al., 2021; Baldwin et al., 2022) or Transformers (Sabater et al., 2023; Peng et al., 2023). Some methods combine several frames for analysis instead of using them one by one. Typically, results for a set of frames are obtained by aggregating the intermediate results of each frame with RNNs (Innocenti et al., 2021; Weng et al., 2021), CNNs (Amir et al., 2017; Innocenti et al., 2021), or attention-based methods (Sabater et al., 2023; Zhang et al., 2022; Wang et al., 2022). Differently, a recent approach introduces a novel event representation, Group Token, to train an event-based ViT backbone called Group Event Transformer (GET) (Peng et al., 2023). In our pipeline, we build dense event representations with FIFO queues, similar to Sabater et al. (2023), and we explore the aggregation of several dense frame predictions during inference.

Table 1 positions the new data released in this work, EventSleep2-data, with respect to a summary of relevant public benchmarks for action recognition using event camera data. Most of them target general daily actions, while ASL-DVS (Bi et al., 2020) and SL-Animals-DVS (Vasudevan et al., 2021) are specialized for sign language gesture recognition. EventSleep is the only one focused on action recognition in dark environments in the context of medical assistance applications.

2.3. Sleep activity recognition with event cameras

The first exploration of sleep activity recognition using event cameras, in the context of health studies, was introduced in EventSleep (Plou et al., 2024). That work served as a proof-of-concept to assess the potential of this setup for supporting medical studies on sleep disorders. EventSleep proposed a benchmark dataset containing short, simulated sleep movements recorded under near-darkness conditions with an event camera and an infrared reference camera. Participants performed ten predefined actions, including six movement types and four static postures, designed to mimic behaviors relevant to sleep disorder analysis. The recordings were captured from a top-down view under different illumination and covering configurations, providing a controlled yet challenging environment with high noise and subtle motion cues.

To address these challenges, the EventSleep baseline employed event frame representations based on EvT+ (Sabater et al., 2023) and evaluated different backbones, with a ResNet-18 (He et al., 2016) trained on two-channel event frames yielding the best performance (hereafter EventSleep1-net). Bayesian calibration strategies, including Laplace ensembles (Eschenhagen et al., 2021), were further incorporated to improve prediction reliability in medical contexts. Although this work demonstrated the feasibility of event-based sleep monitoring, it was limited to short, pre-segmented recordings under controlled recording conditions. The present work extends this framework toward full-night, real-world monitoring through a new dataset (EventSleep2-data) and updated recognition model (EventSleep2-net).

3. EventSleep2-data

The original EventSleep dataset (EventSleep1-data) was limited to short, controlled recordings where subjects simulated isolated movements under laboratory conditions. While this setting allowed the initial exploration of event-based sensing for sleep activity recognition, it lacked the variability, continuity, and scale required for realistic applications. EventSleep2-data extends this benchmark, providing researchers with a novel and unique resource for sleep activity recognition and analysis. It introduces complete-night recordings collected in natural home environments, captured through a motion-triggered acquisition pipeline that stores only relevant motion clips while discarding inactive periods. This design makes efficiently recording full nights (7–8 h) possible, and facilitates studies of real-world sleep behavior at scale. Additionally, several nights include synchronized EEG measurements, enabling multimodal investigations that link visual motion cues with physiological sleep signals.

Setup and scene configuration. The event camera data recorded in this dataset was obtained through a DVXplorer camera with 640px by 480px resolution. The camera was attached to a metallic support next to an infrared spotlight, which provides visibility to the camera without perturbing the subjects' sleep, allowing us to set the sensitivity to *high* instead of *very high*. The metallic support is put on top of a tripod facing down at a -60° angle, which is set to a 2 m height above the floor as shown in Fig. 2. The camera and spotlight are connected to a laptop placed outside the bedroom through USB extenders. Additionally, nights that include EEG also use a 2-channel neuroheadband López-Larraz et al. (2023) that is connected to an independent tablet-like device wirelessly (See Fig. 3). All nights were recorded in personal bedrooms arranged freely by the subjects, with no layout restrictions other than the tripod placement, which must to the left side of the bed and placed at a far enough distance so that the bed is entirely within the field of view of the camera. Bed sizes range from 1.05 to 1.40 m wide and 1.90 to 2.00 m long, and bedclothes vary from duvets to light blankets. Lastly, all nights were recorded in as much darkness as possible.

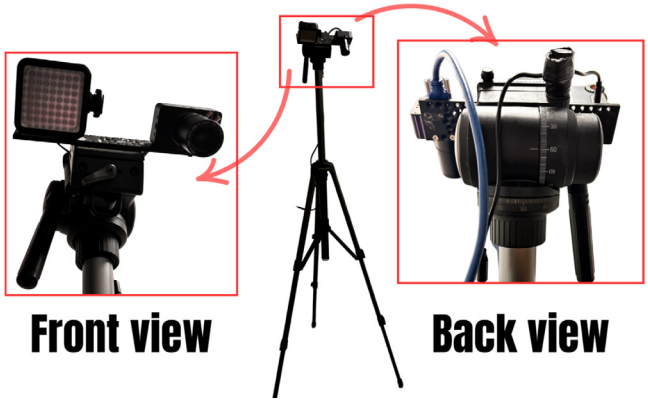

Fig. 2. EventSleep2-data recording scenario. The event camera and infrared spotlight are shown in the front view. The back view shows the camera's inclination (60 degrees).

Recording procedure. The setup described above is mounted in the bedroom of each subject during different nights. Unlike previous works (Mohammadi et al., 2020; Carmona et al., 2023), the people portrayed in our recordings are not patients nor part of a clinical trial. The recordings were taken and released with their full knowledge and permission. However, unlike in EventSleep1-data where subjects were acting, subjects are actually sleeping during full-night sessions in their homes. This fundamental difference required a completely new recording procedure. While EventSleep1-data continuously captured events throughout short (3-4 min) sessions—producing large amounts of noisy, low-information data during static periods—EventSleep2-data employs a motion-triggered acquisition pipeline that records only when movement occurs. This strategy drastically reduces storage requirements and enables realistic 7–8 h overnight recordings without compromising relevant motion information.

This triggering mechanism filters event packages recorded at 1 ms resolution, based on the number of events they contain. Specifically, we pre-computed a threshold set at the 95th percentile of the number of events observed in all event packages captured during a 10-second baseline recording of the scene without motion. To reduce the influence of spurious event peaks, the number of events per package is averaged over a 0.2-second sliding window. Thus, an event package is stored, both raw and filtered, whenever this average exceeds the threshold. The threshold is recalculated hourly to account for possible illumination changes. The first event package of each night is always recorded for synchronization purposes, whether it exceeds the average or not. The duration of the event recording is set to a maximum of 7 h, though the subjects are allowed to finish earlier if they wish.

Additionally, we incorporate a lightweight neuroheadband to complement the event recordings with synchronized brain activity signals. The neuroheadband is a specialized wearable electroencephalography (EEG) device designed for unobtrusive and reliable monitoring of brain activity during sleep, particularly suited for research in real-world settings and clinical studies. This system prioritizes user comfort and ease of use (Gallego et al., 2024), facilitating long-term sleep data collection outside of traditional laboratory environments. The headband key features include its textile-based headband design, which is intended to enhance subject comfort during overnight recordings. This device uses pre-gelled snap-on sensors to acquire signals. These sensors are strategically positioned on the front of the scalp to capture the relevant EEG activity necessary for sleep analysis.

Challenges. The presented dataset poses several interesting challenges for related research. (1) Events are not recorded while the subject is still, so how do we predict the static labels (e.g., *lie up*)? (2) Actions are real, not enacted, meaning several actions can happen at the same

Fig. 3. EEG wearable device. To the left, the headband is placed between the amplifier (top) and the sensors. The tablet device (right) is used to record the data.

time, or an action outside the pre-defined labels may occur. Is zero-shot action recognition necessary? (3) Ideally, the sleep activity analysis should be performed online to allow for interventions during sleep to help with the subject's pathologies. This brings an extra challenge due to the **computational requirements** to run the recognition tasks online.

Labels. The labels correspond to 10 classes, which are the same as those included in the EventSleep-1 dataset. Six of them are motion activities: (0) HeadMove, (1) Hands2Face, (2) RollLeft, (3) RollRight, (4) LegsShake, (5) ArmsShake, and the remaining four are static activities: (6) LieLeft, (7) LieRight, (8) LieUp, and (9) LieDown (See Fig. 4). The selection of these activities is motivated by the combination of simple but significant information needed by the experts analyzing the data to diagnose common sleep disorders. The four static activities (classes 6 to 9) correspond to the standard sleep poses used in medical analysis (left, right, supine, and prone, respectively). Classes 2 and 3 enhance the standard ones by providing information about transitions among them. Classes 4 and 5 may help diagnose periodic limb movement of sleep disorder (PLMS) with no additional sensors such as EMG or IMU attached to the limbs. Finally, classes 0 and 1 are helpful to decode and preprocess the brain signal, as they indicate potential sources of artifacts in the EEG signal caused by head movements or by the user touching the sensor. Motion activity labels were manually annotated by reviewing event reconstructions and marking the start and end frames of each action, following the original dataset's guidelines. Static pose labels were then added by identifying the event reconstruction frames (typically during a 'RollLeft' or 'RollRight' motion label) where the subject transitioned to a new position.

These frame-level annotations were then mapped to the time domain, allowing their use for EEG-based action recognition, as both modalities are temporally aligned. Static labels are extended across time segments without event frames, covering periods of inactivity. Thus, time segments containing events—triggered by the subject's activity—include both motion and static pose labels, whereas segments without events contain only a static pose label.

Content and accessibility. EventSleep2-data comprises 10 full-night recordings, totaling approximately 70 h of sleep monitoring from three different subjects. Six of these recordings include synchronized EEG data. All data and metadata will be made publicly available upon acceptance via Synapse, as a new set of data within the current EventSleepdatarepository.

EEG discussion. A total of six out of ten nights include EEG signal captured with the neuroheadband setup. The dataset extension includes nights with and without the headband device to avoid biases in movement (e.g., scratching or relocating the band). The headband captures a two-channel signal through electrode patches placed on the sides of the subjects' foreheads, just above the eyebrows. The signal metadata includes UTC annotations for synchronization with the event pipeline.

Although the approach presented in this paper does not use the captured signal for action recognition, we make it available for future works to use. We believe the correlation between movement and EEG can be helpful for research in both fields. Example applications range from tasks typically performed over EEG, such as sleep stage prediction, which could benefit from action recognition to match each sleep stage with plausible movements, to enriching further activity recognition approaches with EEG data.

We provide a preliminary qualitative study on how EEG and event data correlate through several examples extracted from our recordings. After observing an aligned display of the number of events and the EEG signal, as portrayed in Fig. 5, it is noticeable that peaks in events often correspond with high amplitude regions of the EEG. This particular finding may help identify movement-related signals, or avoid error cases, product from hardware inaccuracies caused by movement.

4. EventSleep2-net

The original EventSleep1-net was designed to process continuous recordings in which both motion and static periods contained event activity. However, this design is not compatible with the motion-triggered format of EventSleep2-data, where static clips contain no events. To address this limitation, the proposed EventSleep2-net introduces a new architecture specifically adapted to the recording strategy of EventSleep2-data, enabling accurate prediction of both motion and static poses throughout complete-night sequences.

4.1. Events representation

As in our previous work (Plou et al., 2024), we adopt the EvT+ (Sabater et al., 2023) dense frame-based representation, where asynchronous events are aggregated into fixed-rate frames. In other words, we build frames $\mathbf{F} \in \mathbb{R}^{H\times W\times 2}$ at frame rate Δt. Each pixel stores the timestamp of the most recent event that took place in that region within a temporal window of length T_M, normalized to the range $[0, 1]$. Pixels without recent events are set to `NaN`.

This formulation is identical to EventSleep1-net and ensures comparability across datasets. The only addition here is the distinction between the clip-based and full-sequence data formats illustrated in Fig. 6. In the clip-based setting, events are processed independently per action, while in the full-sequence setting, temporal windows are continuous, allowing overlap between actions and better reflecting realistic acquisition.

4.2. Base architecture

The proposed EventSleep2-net architecture, illustrated in Fig. 1, builds upon the EventSleep1-net design (Plou et al., 2024) but introduces a new dual-head configuration specifically adapted to motion-triggered recordings. The base model uses a shared feature extraction backbone (**Encoder**) coupled with specialized classification heads (**Motion Head** and **Static Head**) to effectively process event data for human activity and posture recognition. This design choice is motivated by the hypothesis that both dynamic movements and static postures occur at the same time. Despite their distinct nature, they exhibit shared spatial features that can be efficiently learned by a common network trunk.

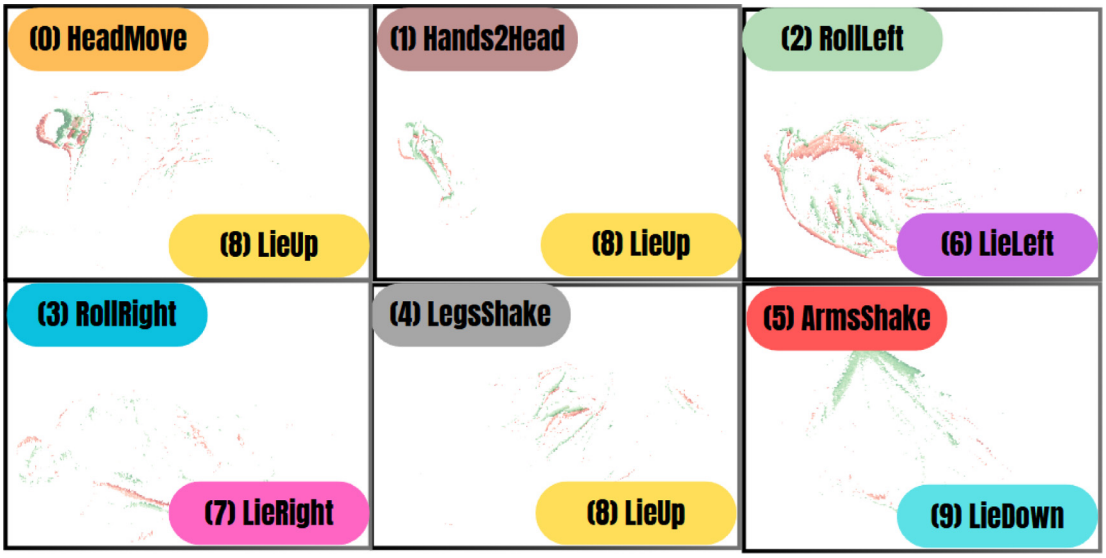

Fig. 4. Examples of EventSleep2-data event frame annotations. Unlike the original EventSleep1-data, where each event frame had a single label, EventSleep2-data assigns two labels per frame: a **motion label** (shown on the upper-left side of the frame) and a **static pose label** (on the lower-right side). Static labels are extended across time segments without event frames, covering periods of inactivity.

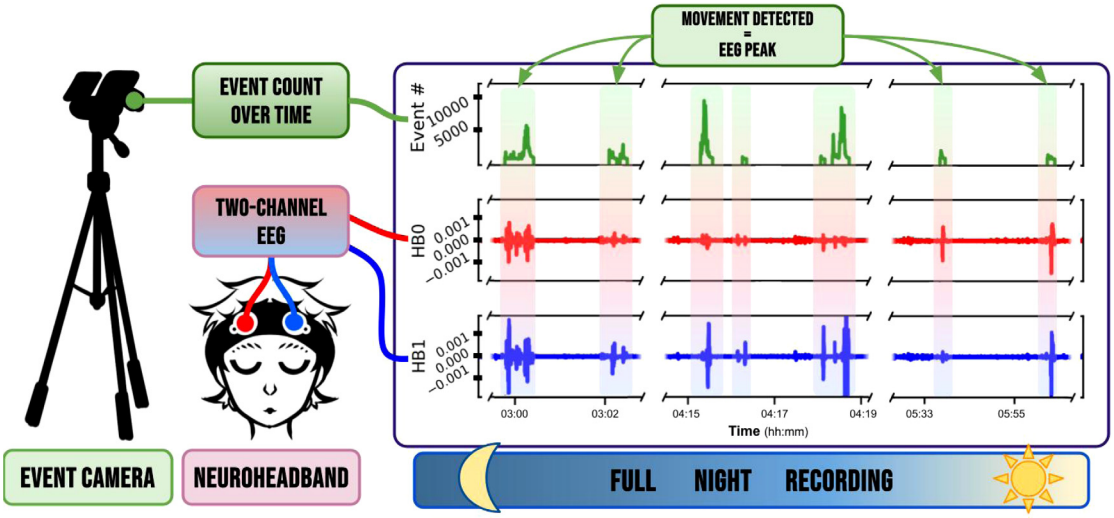

Fig. 5. Synchronized EEG and event data. Plot for the event count (top, green) and the two-channel EEG (mid and bottom, red and black) over time for a sample from the EventSleep2-data training set. The event count plot shows peaks when movement is detected and zero otherwise. The EEG plot shows the measured brainwave amplitude. Notice how detected movements (in shaded background) are often coincidental with peaks in at least one of the recorded channels. Movements last a few seconds, while the waits between them can last for hours. X axis is broken for visualization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

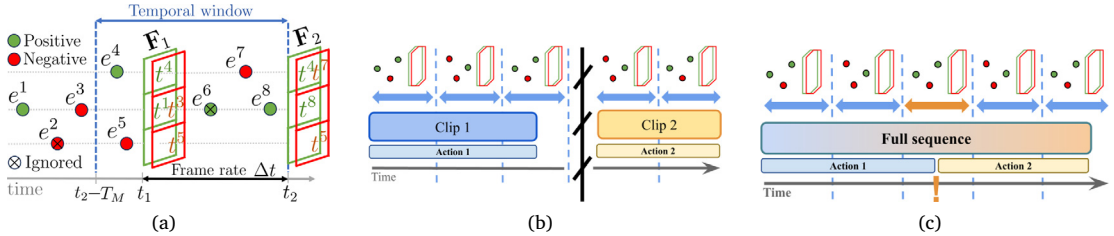

Fig. 6. Event frame construction process. (a) Each Δt an event frame $\mathbf{F}$ is built. It accumulates relevant event information from the defined preceding *Temporal window* of length T_M. It stores the timestamp from the most recent positive and negative event that occurred for each pixel position within that window. If none occurred, that pixel is "empty" (like a "NaN" value). Event frame construction can differ depending on whether (b) independent clip durations are known, so frames are padded with zeros between clips if needed or (c) in a more realistic setting, the algorithm processes a continuous stream of events, potentially mixing events from two different actions, and offsetting the temporal window.

4.2.1. Encoder

The encoder is identical to the one used in EventSleep1-net and is based on a ResNet-18 (He et al., 2016) backbone pretrained on ImageNet (Deng et al., 2009) and fine-tuned on event frame representations. This backbone efficiently extracts spatial features from the two-channel event frames described in Section 4.1.

4.2.2. Specialized classification heads

The main novelty of EventSleep2-net lies in its dual-head prediction scheme. The feature vector produced by the encoder is duplicated and fed to two separate classification heads. This design is intended to address the objective of concurrently predicting both the movement and the static pose of the subject. This dual-head design is motivated by the need to predict both aspects concurrently from the same input features. The *Motion Classification Head* is optimized to recognize patterns indicative of movement, while the *Static Classification Head* focuses on patterns defining static poses. By separating these tasks, each head can optimize its final layers to better interpret the shared features specifically for either dynamic patterns (actions) or static configurations (positions), potentially improving the accuracy of both predictions.

Motion classification head. This head has been developed for the specific purpose of classifying activities that are distinguished by substantial body movement. The feature vector is received from the shared backbone, and it is processed through a sequence of fully connected layers. The purpose of these dedicated layers is to learn the specific mappings from the shared spatial features to the high-level concepts of dynamic actions.

Static classification head. This head focuses on identifying static postures. Similar to the movement head, it also takes the identical feature vector from the shared backbone as input and employs fully connected layers. By having a separate head, the model can specialize in discerning the spatial patterns present in the event data during periods of relative stillness, which might be obscured or processed differently if combined directly with high-motion classes.

4.3. Training: two-step training on event data

Our encoder is initialized with ImageNet (Deng et al., 2009) pretrained weights, which provide generic spatial representations. Then, the training strategy of EventSleep2-net follows a progressive adaptation pipeline to bridge the domain gap between standard visual features and motion-triggered event data.

Step 1: Base training on EventSleep1-data. As a first step, we train the EventSleep2-net using only EventSleep1-data, where continuous sequences are employed instead of the short pre-segmented clips used in the original setup. Static clips are replaced with empty event windows to simulate the motion-triggered format of EventSleep2-data. This step prepares the model to handle full-sequence recordings in which static periods contain no event information.

Step 2: Fine-tuning on EventSleep2-data. To finalize the training, a second fine-tuning phase is performed using the validation sequences of EventSleep2-data to adapt the model to the real-world recordings acquired in home environments. This step compensates for differences between the controlled laboratory data of EventSleep1 and the more natural conditions of EventSleep2, which exhibit lower event density and higher variability. Three sequences are used for fine-tuning and one unseen sequence for validation.

The network parameters are optimized by minimizing a composite loss function $\mathcal{L}$ that combines the motion and static classification losses:

$$\mathcal{L} = \alpha \mathcal{L}_m + \beta \mathcal{L}_s, \tag{1}$$

where $\mathcal{L}_m$ and $\mathcal{L}_s$ denote the cross-entropy losses for motion and static classes, respectively. The weighting coefficients $\alpha = 0.6$ and $\beta = 0.4$ reflect the proportion of motion and static labels within the dataset. Training is conducted for 50 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 32. The model achieving the highest validation accuracy is selected for final evaluation.

4.4. Inference

During inference, the network processes continuous event streams frame by frame. The Motion Head produces per-frame action probabilities, while the Static Head predicts the most likely posture. When no events are detected—i.e., during inactive periods—the last predicted static pose is propagated until new motion is detected. This propagation mechanism allows the model to maintain consistent pose predictions throughout full-night recordings, even in the absence of event activity, effectively bridging the gaps between motion-triggered clips.

4.5. Bayesian deep learning: Laplace ensembles module

Following our previous work (Plou et al., 2024), we also incorporate Bayesian Deep Learning methods to improve prediction calibration and reliability. Specifically, we use Laplace Ensembles (Eschenhagen et al., 2021), which combine deep ensembles (Lakshminarayanan et al., 2017) with Laplace approximations (Daxberger et al., 2021) applied to the last layer of the network. This configuration, identical to that used in EventSleep1-net, enables well-calibrated probabilistic predictions and provides robustness to noisy event data.

5. Experiments

The following experiments evaluate the proposed EventSleep2-net approach for sleep activity recognition on bothEventSleep1-data and EventSleep2-data, and illustrate the open challenges of this application. For clarity, EventSleep1-net and EventSleep2-net will be referred to as EvS1-net and EvS2-net.

5.1. Experimental setup

Baselines. We compare our current work with *Random guess* as baseline, GET (Peng et al., 2023) as the state-of-the-art for event activity recognition and ResNet-E (Plou et al., 2024), renamed in this work as EvS1-net for clarity, as the current state-of-the-art for sleep activity recognition with event cameras.

Pre-processing. Our pre-processing of the event data captured involves constructing event frames. Our event frames are constructed every $\Delta t = 0.15$s with $T_M = 0.512$s, to capture enough temporal information to classify the action. We also reduce the resolution by a half, obtaining frames of grid resolution 320×240.

Training process. The model was trained using the two-step process detailed in Section 4.3 and optimized with the standard cross-entropy loss. For Laplace-Ensembles (LE) we use 5 ensembles and we fit Laplace approximation to the last layer of both heads.

Step 1: Training on EventSleep1-data. We use the original training and validation sets for training and to adjust hyper-parameters, respectively. In order to train the two heads, we extend the labels of EventSleep1-data assigning to each action the corresponding static position. We ignore the event data from static clips. Training was conducted for 25 epochs. Adam optimizer was employed with a learning rate of $1e-4$, and a batch size of 32.

Step 2: Fine-tuning on EventSleep2-data validation samples. It was conducted for 50 epochs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 32. The loss weights (Eq. (1)) were set to $\alpha = 0.6$ and $\beta = 0.4$, reflecting the proportion of motion and static samples in the dataset. The checkpoint achieving the highest validation accuracy was selected for evaluation.

Metrics. We evaluate classification performance with the Accuracy, defined as the percentage of samples for which the top predicted class matches the ground truth label.

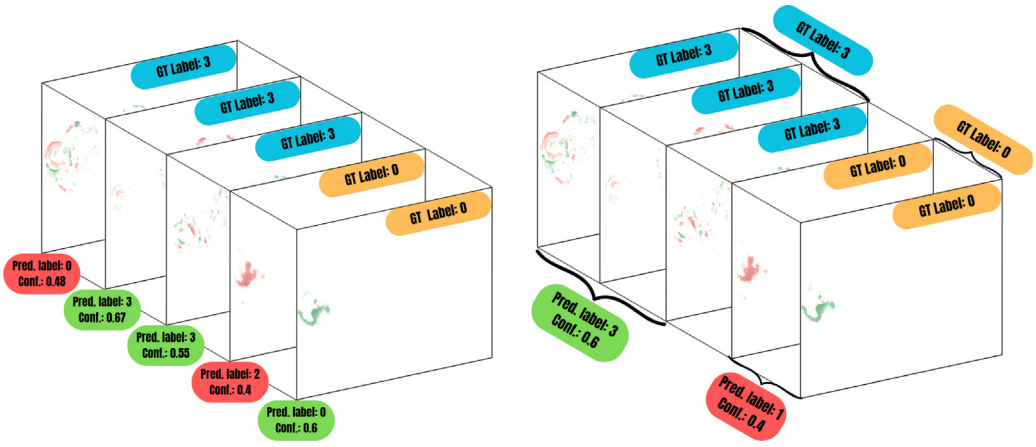

Fig. 7. Illustration of evaluation strategy **per frame** (a) and **per clip** (b), as explained in Section 5.3. This example, corresponding to the *Motion head* output, shows for each frame: ground truth (GT) label; predicted label (Pred. Label); prediction confidence (Conf.) level.

5.2. Results on EventSleep2-data

EventSleep2-data extends the original EventSleep1-data with more challenging and realistic recordings of people sleeping during the whole night (See Section 3). An important difference in these new data recordings is that the system only stores data during clear subject movements, and it stops recording completely when the movement is finished, i.e., when the recording system sees such a low amount (or none) events and interprets that there is no subject activity anymore.

Thus, static clips contain no event data. Consequently, GET and EvS1-net architectures are fundamentally incapable of providing predictions for static clips in EventSleep2-data. As single-output classifiers, these methods cannot propagate static pose predictions from motion clips; for any given motion frame or clip, they will predict a motion label rather than a static one.

Table 2 shows the results on EventSleep2-data. We provide two ways of evaluating the predictions: **per frame**, i.e., assign and evaluate separately a label for each frame, and **per clip**, i.e., assign a label to a group (clip) of frames that are known to form a single action (See Fig. 7).

To ensure a fair comparison with SOTA, we will evaluate our EvS2-net using two different configurations. The first, designated EvS2-net (base), represents the most comparable setup to SOTA, specifically by omitting three key components: the Step 2 fine-tuning (Section 4.3), and the Laplace Ensembles module. Our full, optimal configuration, which includes all these modules, is referred to simply as EvS2-net. The results show a clear advantage of the new presented method to process continuous, real-world recording data.

5.3. Results on original EventSleep1-data

To provide a comprehensive performance context, especially since most baselines are not fully evaluable on the new EventSleep2-data, we extend our comparison to the original EventSleep1-data. Specifically, we focus on comparing the current state-of-the-art (SOTA) on this benchmark, EvS1-net, against our EvS2-net (base).

Clip data format experiment. Originally, EventSleep1-data was only evaluated with *Clip data format*. This configuration takes the event data split in independent Clips, that contain a single action each, and constructs the event frames on each Clip independently. Table 3 first compares the performance of our approach with respect to the current SOTA in the test split of EventSleep1-data. Maintaining the exact configuration and set up of the original work, we obtain similar performance for the Motion labels.

It is important to note that EvS2-net is not able to make predictions for static classes when using the clip-based formatting, where clips are processed isolated. It can only predict static classes when handling full sequences, by predicting the static pose corresponding to the action clip, i.e., the end pose after the motion, and propagating this pose through the following static segment without events. Hence the 'N/A' in Table 3.

Full-sequence data format experiment. The previous experiment puts our current work in context with the original EventSleep1-data. However, our current goal is to handle continuous, real-world recordings. Towards this goal, we run a second experiment, summarized in second half of Table 3, using the full sequences of the EventSleep1-data test set. As previously mentioned, EvS2-net is specifically designed to operate on realistic full sequences recorded with the motion-triggered procedure. When working with full sequences it is able to use the motion-triggered procedure to distinguish motion from static clips, ignoring all the information from the static clips, predict the static pose at the end of each motion clip, and propagate this pose during the following inactive segments. This leads to a clear improvement on static clips and makes it more suitable for real-world scenarios.

6. Configuration studies

6.1. Ablation study: from EvS2-net to EvS2-net (base)

Table 4 presents an ablation study of our EvS2-net architecture conducted on the EventSleep2-data. This analysis summarizes the impact of the core modules, Fine-Tuning step (FT) and the Laplace Ensembles (LE), on the final prediction accuracy and computational time.

The ablation study confirms the impact of each major architectural element. Fine-tuning (FT) provides an essential base improvement by adapting the model to real-world data, while the subsequent incorporation of the Laplace Ensembles (LE) module significantly enhances prediction overall accuracy. Including LE brings an expected $x5$ increase on the inference time and memory requirements (note LE uses 5 ensembles). Note this is still sufficiently fast for the application requirements (as discussed in Section 6.3) and the approach confidence calibration is significantly improved, as illustrated in the following experiment.

Table 2
Results on EventSleep2-data test set. Accuracy for related baselines and our EvS2-net with its best and base configuration (trained on similar conditions than the baselines). Event frames are always computed on complete sequences (*Full Sequence* data format), as described in Section 4.1. Accuracy is reported Per Frame (individual frame predictions) and Per Clip (aggregated predictions over the clip). Note that only EvS2-net can be fine-tuned with EventSleep2-data, since it is the only approach suited to handle the empty static-class clips. Best value in bold.

	Accuracy - Per Frame			Accuracy - Per GT Clip		
	Motion Labels	Static Labels	AVG	Motion Labels	Static Labels	AVG
Methods trained on EventSleep1-data **without** fine-tuning on EventSleep2-data						
Random guess	0.17	0.25	0.20	0.17	0.25	0.20
GET (Peng et al., 2023)	0.24	N/A	–	0.23	N/A	–
EvS1-net (Plou et al., 2024)	0.31	N/A	–	0.35	N/A	–
EvS2-net (base)	0.35	0.33	0.34	0.38	0.33	0.36
Methods trained on EventSleep1-data **with** fine-tuning on EventSleep2-data						
EvS2-net (ours)	**0.46**	**0.43**	**0.45**	**0.50**	**0.51**	**0.50**

N/A: Not evaluable with this method.

Table 3
Results on EventSleep1-data test set: EvS1-net vs EvS2-net (base). We build the event frames on manually separated clips (*Clip*) as reported in the previous work by Plou et al. (2024), or on complete sequences (*Full Sequence*), as described in Section 4.1. Accuracy is reported Per Frame (individual frame predictions) and Per Clip (aggregated predictions over the clip). Best value in bold.

	Accuracy - Per Frame			Accuracy - Per GT Clip		
	Motion Labels	Static Labels	AVG	Motion Labels	Static Labels	AVG
Clip input data format						
EvS1-net	0.77	0.49	0.66	**0.97**	0.61	0.82
EvS2-net (base)	**0.79**	N/A	–	0.95	N/A	–
Full Sequence input data format						
EvS1-net	0.77	0.52	0.67	0.97	0.62	0.83
EvS2-net (base)	**0.78**	**0.57**	**0.69**	**0.99**	**0.92**	**0.96**

N/A: Not evaluable with this method.

Table 4
EvS2-net ablation study: From the base configuration to our final configuration (last row). We report inference time (Inf. T.), parameters required (Num. Params.) and accuracy on the test samples of the EventSleep2-data. We analyze the impact of our modules fine-tuning (FT) and Laplace Ensembles (LE).

EvS2-net			Accuracy - Per Frame			Accuracy - Per GT Clip		
Config.	Num. Params.	Inf. T. (ms)	Motion Labels	Static Labels	AVG	Motion Labels	Static Labels	AVG
Base	11.7M	0.46	0.35	0.33	0.34	0.38	0.33	0.36
Base + FT	11.7M	0.46	0.36	0.38	0.37	0.45	0.44	0.44
Base + FT + LE	58.7M	2.55	**0.46**	**0.43**	**0.45**	**0.50**	**0.51**	**0.50**

6.2. Confidence calibration

Confidence calibration is the measure of correlation between a model's predicted confidence and its actual accuracy. This correlation is crucial for safety-critical applications, such as medical diagnosis, where accurately interpreting the model's certainty is essential.

To assess the calibration of our predictions, we employ set of metrics based on the histogram of Reliability Diagrams as proposed in DeGroot and Fienberg (1983). These histograms illustrate the discrepancy between predicted confidence and observed accuracy across a set of bins. We divide the prediction outputs into $M = 10$ bins based on their confidence scores. For each bin, we compute the average confidence C_m and accuracy A_m. Based on these statistics, we compute two metrics commonly used in the literature (Neumann et al., 2018): the Average Calibration Error (ACE) defined as follows,

$$\mathrm{ACE} = \frac{1}{M^+} \sum_{m=1}^{M^+} \left| C_m - A_m \right|, \tag{2}$$

and the Maximum Calibration Error (MCE),

$$\mathrm{MCE} = \max_m \left| C_m - A_m \right|, \tag{3}$$

where M^+ is the number of bins containing samples.

Table 5
EvS2-net ablation study: Effect of Laplace Ensembles module on the confidence calibration. We report calibration metrics (MCE, ACE) on the validation samples of the EventSleep2-data, separately for Motion and Static labels.

EvS2-net Config.	Motion calibration metrics		Static calibration metrics	
	MCE ↓	ACE ↓	MCE ↓	ACE ↓
Base	0.55	0.28	0.66	0.31
Base + FT	0.52	0.31	0.55	0.27
Base + FT + LE (ours)	**0.27**	**0.14**	**0.41**	**0.13**

Table 5 summarizes the calibration results. These metrics clearly highlight the impact of the Laplace Ensembles module, demonstrating its effectiveness not only in improving raw accuracy (Section 6.1) but also in significantly enhancing confidence calibration.

6.3. Complete pipeline discussion

The presented ablation study in this section has systematically evaluated the contributions of distinct components incorporated to our base EvS2-net approach. The results clearly show that applying the fine-tuning procedure and Laplace ensembles together yields the most substantial improvement in accuracy and calibration metrics (See Tables 4 and 5). Among other alternative studied, we include in Appendix B an experiment that incorporates a sliding window to smooth the predictions with recent previous event frames processed. The results obtained suggest that the robustness in the predictions obtained thanks to the LE module are sufficient and the sliding window does not add benefit to the overall approach. As discussed next, the performance obtained is promising for applicability in real scenarios, although it presents several intrinsic limitations to be tackled in future work.

The performance on the different approaches in our new EvS2-data benchmark is affected by certain **common limitations**. First, while there is room for improvement in overall accuracy, the current evaluation suffers from noise introduced by the difficulty of annotating each activity occurrence very accurately. Second, real-world recordings

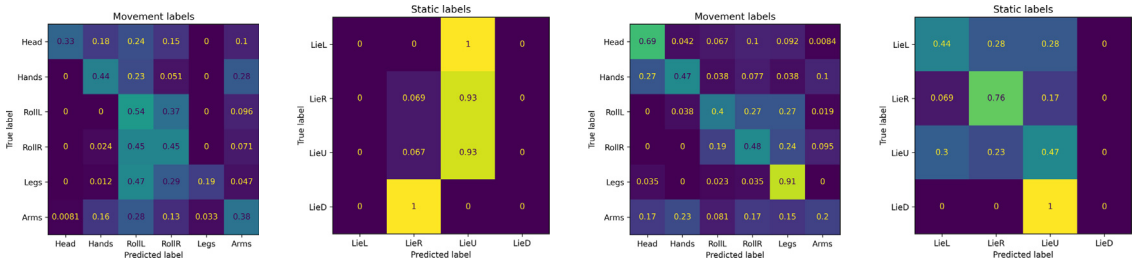

Fig. 8. Confusion matrices for motion labels and static labels using *EvS2-net-base* (a–b) and using *EvS2-net* (c–d) in *EventSleep2-data*. The model visually leans towards the 'RollLeft' and 'LieUp' classes, while ignoring 'Arms' and 'LieDown', likely due to class imbalance. Real night recordings are limited and not directed, so class coverage cannot be controlled.

frequently contain clips that combine multiple simultaneous movements (e.g., arm and leg motion during a roll), which complicates the assignment of a single activity label. Third, static labels are affected by class imbalance. These factors, as revealed in the confusion matrices (Fig. 8) and qualitative results (Fig. 9), contribute to misclassifications. For future work, the current single-class prediction design could be expanded into a multi-class architecture where the network can predict simultaneous activity labels to better account for complex, compound movements. Additionally, extending the dataset with a significantly larger volume of real-world recordings remains a highly relevant next step toward applicability in real scenarios.

Regarding **applicability** in real-world settings, we observe a good performance of the proposed approach for the envisioned scenario (automated reports that can provide useful information to experts shortly after the sleep period has finished). Our pipeline allows us to process the event-camera data as it arrives. Our implementation constructs a frame every 0.15s (on average, matching our selected Δt). Construction time scales linearly ($\mathcal{O}(n)$) with the number of events in the time window. Combined with the short inference times and the motion-triggered capture system, we can store and process event frames either in parallel or during inactivity periods.

Overall, EventSleep2 offers a significant advancement over prior work, primarily by enabling the processing of full-night, real-world recordings. Our EvS2-net is uniquely designed to handle the motion-triggered nature of the new full-night recordings.

7. Conclusion

This work highlights the potential of event-based vision as an effective and low-intrusion solution for in-home sleep activity recognition. The introduction of EventSleep2-data—a substantial extension of the original dataset with full-night recordings under real-world conditions—marks a significant step towards practical applications of this technology. The inclusion of synchronized EEG for selected sessions adds further depth and relevance to the dataset, enabling future exploration of multimodal correlations.

We also proposed EventSleep2-net a novel dual-head recognition architecture capable of jointly classifying motion events and static sleep poses. This approach proves more robust than previous baselines when handling the inherent sparsity of overnight data. Unlike the original EventSleep framework, which was limited to short, controlled sequences, this new version enables continuous overnight analysis based on a motion-triggered acquisition and an adapted inference strategy. Our experiments on both EventSleep1-data and EventSleep2-data demonstrate state-of-the-art performance, supported by detailed ablation studies that validate our design choices. Overall, this work consolidates a complete and scalable framework—from data collection to model design—capable of supporting real-world, long-term sleep monitoring scenarios. Together, the dataset and proposed framework advance the role of event-based cameras as a key computer vision technology to assist in the sleep research domain, and pave the way for more scalable, non-invasive sleep monitoring solutions. Future work includes the design of a robust computing platform for clinical validation of EventSleep2-net.

CRediT authorship contribution statement

Nerea Gallego: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Carlos Plou:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Miguel Marcos:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation, Conceptualization. **Pablo Urcola:** Validation, Resources, Funding acquisition. **Luis Montesano:** Validation, Resources, Funding acquisition. **Eduardo Montijano:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition. **Ruben Martinez-Cantin:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition. **Ana C. Murillo:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by PID2024-159284NB-I00, PID2021-125514NB-I00, PID2024-158322OB-I00, PID2021-125209OB-I00, and AIA2025-1635 grants funded by MCIN/AEI/10.13039/501100011033 ERDF/NextGenerationEU/PRTR, grant no. 101135782 (MANOLO project) funded by the European Union, two DGA scholarships and project T45_23R.

Appendix A. Additional dataset details

Participants and Ethical considerations. The participants present in the dataset extension are three adults with no sleep-related medical conditions. All participants are in their mid twenties and include one female and two males. All three participants were volunteers and gave their consent signing a written form, which included instructions, an explanation of the experiment and authorization to publicly release the recordings following the General Data Protection Regulation (GDPR), one of the most restrictive regulations in this matter.

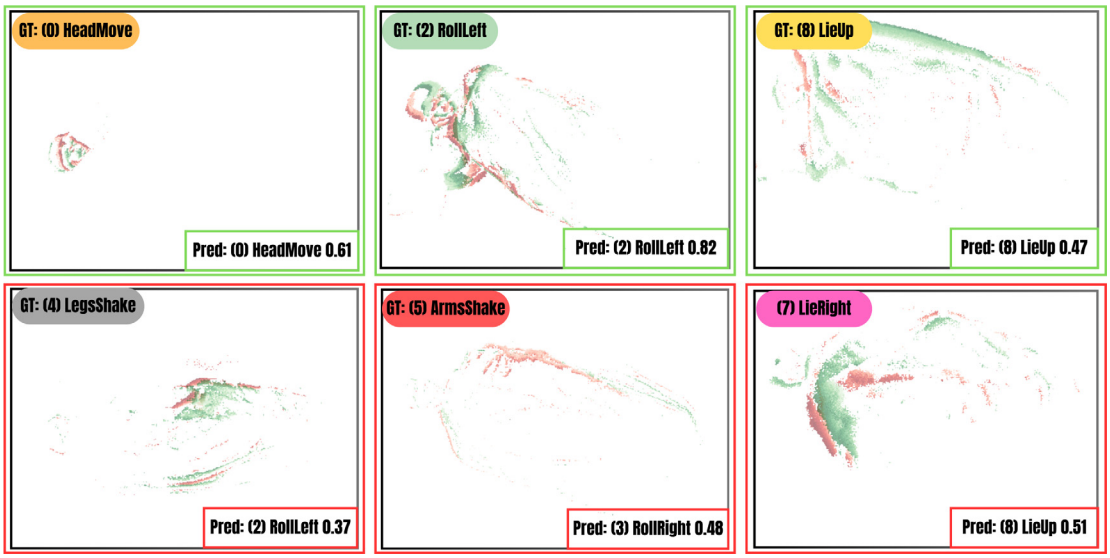

Fig. 9. Qualitative results of EvS2-net using EventSleep2-data. The upper row shows actions that the model classifies correctly, while the lower row shows examples of misclassifications. Static labels correspond to the state after the portrayed frame. Actions performed during sleep can be confusing due to subtle or ambiguous movements. People also tend to perform multiple actions at once, such as moving their arms and legs while rolling (see bottom-left and middle). Static labels are mostly confused due to class imbalance, often predicting 'LieUp'.

Table B.6
EvS2-net ablation study: effect of window size (Size) and sliding stride (Stride). Results are the average accuracy in all validation sequences from EventSleep2-data.

Size/Stride	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.84	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2	0.83	0.83	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
3	0.81	0.82	0.82	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
4	0.80	0.82	0.82	0.81	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
5	0.78	0.80	0.82	0.82	0.81	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
6	0.77	0.79	0.80	0.81	0.81	0.81	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
7	0.76	0.77	0.78	0.80	0.81	0.80	0.80	N/A	N/A	N/A	N/A	N/A	N/A	N/A
8	0.75	0.77	0.78	0.79	0.80	0.80	0.80	0.79	N/A	N/A	N/A	N/A	N/A	N/A
9	0.73	0.75	0.77	0.78	0.78	0.79	0.79	0.79	0.78	N/A	N/A	N/A	N/A	N/A
10	0.72	0.74	0.76	0.77	0.77	0.79	0.78	0.79	0.78	0.78	N/A	N/A	N/A	N/A
11	0.71	0.73	0.74	0.75	0.77	0.78	0.78	0.78	0.79	0.79	0.77	N/A	N/A	N/A
12	0.70	0.72	0.73	0.74	0.75	0.77	0.77	0.79	0.78	0.78	0.77	0.76	N/A	N/A
13	0.68	0.70	0.72	0.73	0.74	0.76	0.77	0.78	0.78	0.78	0.77	0.76	0.76	N/A
14	0.68	0.69	0.71	0.73	0.74	0.74	0.76	0.78	0.77	0.78	0.78	0.75	0.77	0.74
15	0.67	0.67	0.69	0.71	0.72	0.72	0.75	0.76	0.76	0.77	0.78	0.74	0.76	0.75

Hardware used. The recording setup included a laptop equipped with a Intel Core™ i7-6700HQ CPU with 8 cores and a NVIDIA GeForce GTX 970M GPU. Model training and testing, and the sliding-window study presented in Appendix B, were conducted using a AMD® Ryzen 9 9950x processor with 16 cores, and NVIDIA GeForce RTX 4090 GPU. Reported times for event frame construction were obtained in a 12th Gen Intel® Core™ i7-12700K CPU.

Appendix B. Additional experiments: Sliding window

In this appendix, we present an additional experiment conducted to assess the impact of a sliding window (SW) mechanism on temporal smoothing and classification stability in the EventSleep2-net framework.

The sliding window module was designed to enhance the temporal coherence of label predictions by leveraging contextual information across consecutive frames. Instead of assigning independent predictions to each frame, the model assigns a single representative label to all frames within a defined window, based on the aggregated prediction confidence.

Given an input sequence $S = \{f_1, f_2, \dots, f_N\}$, where f_t is the frame at timestep t and N is the total number of frames. We define a sliding window of a fixed length W frames. This window moves across the sequence with a stride of S_t frames. Thus, the ith window, encompasses the set frames $\{f_k, f_{k+1}, \dots, f_{k+W-1}\}$, where $k = (i-1)S_t+1$. The label is assigned by taking the maximum value from the summation of softmax applied to the predictions for each frame within the window.

B.1. Experimental setup

We tested sliding-window sizes up to **15 frames** and strides up to **14 frames**. Results on the EvS2-data validation set are reported in Table B.6. The chosen values span the practical temporal scales present in our recordings: the average clip lengths per motion label (in frames) are 11.5, 19.27, 33.67, 33.70, 12.08 and 17.22 for motion labels respectively (std. devs 12.09, 12.63, 20.43, 19.62, 10.53 and 16.16). With the frame generation frequency used in our pipeline ($\Delta t = 0.15s$), these averages correspond approximately to $1.73s$, $2.89s$, $5.05s$, $5.06s$, $1.81s$ and $2.58s$ respectively. These chosen window sizes (up to 15 frames $\approx$ 2.25 s) meaningfully cover the shorter actions while avoiding including entire multi-action segments.

Table B.7
EvS2-net ablation study: comparison of our final configuration without (top) and with (bottom) a sliding window. We report accuracy on the test samples of EventSleep2-data. We analyze the impact of a sliding window (SW) on top of fine-tuning (FT) and Laplace Ensembles (LE).

EvS2-net	Accuracy - Per Frame			Accuracy - Per GT Clip		
Config.	Motion Labels	Static Labels	AVG	Motion Labels	Static Labels	AVG
Base + FT + LE + SW	0.46	**0.44**	0.45	0.47	0.50	0.48
Base + FT + LE (ours)	**0.46**	0.43	**0.45**	**0.50**	**0.51**	**0.50**

B.2. Results and discussion

Table B.6 summarizes the average accuracy for different window and stride configurations across validation sequences. The results show that using a sliding window results in slightly lower classification accuracy than per-frame predictions, with larger windows leading to poorer performance. These results suggest that the model is robust enough and that the smoothing added by the sliding window is unnecessary.

To further assess the impact of this module, Table B.7 compares the performance of the model with and without the sliding window mechanism on the test set. This shows that the sliding window mechanism is not helpful in any of the evaluation modes (per frame or per GT clip). However, these results serve as a reference for future work exploring smarter strategies for long-duration, event-based recordings.

References

Akbarian, S., Ghahjaverestan, N.M., Yadollahi, A., Taati, B., 2021. Noncontact sleep monitoring with infrared video data to estimate sleep apnea severity and distinguish between positional and nonpositional sleep apnea: Model development and experimental validation. J. Med. Internet Res. 23 (11), e26524.

Amir, A., et al., 2017. A low power, fully event-based gesture recognition system. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition.

Baldwin, R., Liu, R., Almatrafi, M.M., Asari, V.K., Hirakawa, K., 2022. Time-ordered recent event (TORE) volumes for event cameras. IEEE Trans. Pattern Anal. Mach. Intell..

Bi, Y., Chadha, A., Abbas, A., Bourtsoulatze, E., Andreopoulos, Y., 2020. Graph-based spatio-temporal feature learning for neuromorphic vision sensing. IEEE Trans. Image Process..

Cai, H., Chen, P., Jin, Y., Zhang, Q., Cheung, T., Ng, C.H., Xiang, Y.-T., Feng, Y., 2024. Prevalence of sleep disturbances in children and adolescents during COVID-19 pandemic: a meta-analysis and systematic review of epidemiological surveys. Transl. Psychiatry 14 (1), 12.

Cannici, M., Ciccone, M., Romanoni, A., Matteucci, M., 2020. A differentiable recurrent surface for asynchronous event-based data. In: European Conf. on Computer Vision.

Carmona, J., Karácsony, T., Cunha, J.P.S., 2023. BlanketSet-a clinical real-world in-bed action recognition and qualitative semi-synchronised motion capture dataset. In: 2023 IEEE 7th Portuguese Meeting on Bioengineering. ENBENG, IEEE, pp. 116–119.

Carter, J.F., Jorge, J., Gibson, O., Tarassenko, L., 2024. Sleepvst: Sleep staging from near-infrared video signals using pre-trained transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12479–12489.

Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., Hennig, P., 2021. Laplace redux-effortless bayesian deep learning. Adv. Neural Inf. Process. Syst. 34, 20089–20103.

DeGroot, M.H., Fienberg, S.E., 1983. The comparison and evaluation of forecasters. J. R. Stat. Soc.: Ser. D (the Statistician) 32 (1–2), 12–22.

Deng, Y., Chen, H., Chen, H., Li, Y., 2021. EV-VGCNN: A voxel graph CNN for event-based object classification. arXiv preprint arXiv:2106.00216.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848.

Dong, Y., Li, Y., Zhao, D., Shen, G., Zeng, Y., 2023. Bullying10k: a large-scale neuromorphic dataset towards privacy-preserving bullying recognition. Adv. Neural Inf. Process. Syst. 36, 1923–1937.

Eschenhagen, R., Daxberger, E., Hennig, P., Kristiadi, A., 2021. Mixtures of Laplace approximations for improved post-hoc uncertainty in deep learning. arXiv preprint arXiv:2111.03577.

Gallego, N., Plou, C., Montesano, L., Murillo, A.C., Montijano, E., 2024. Vision-based feedback on correct sensor placement in medical studies. In: 2024 7th Iberian Robotics Conference. ROBOT, pp. 1–6. http://dx.doi.org/10.1109/ROBOT61475.2024.10796934.

Gao, Y., Lu, J., Li, S., Li, Y., Du, S., 2024. Hypergraph-based multi-view action recognition using event cameras. IEEE Trans. Pattern Anal. Mach. Intell..

Gao, Y., Lu, J., Li, S., Ma, N., Du, S., Li, Y., Dai, Q., 2023. Action recognition and benchmark using event cameras. IEEE Trans. Pattern Anal. Mach. Intell..

Ghosh, R., Gupta, A., Nakagawa, A., Soares, A., Thakor, N., 2019. Spatiotemporal filtering for event-based action recognition. arXiv preprint arXiv:1903.07067.

Han, F., Yang, P., Feng, Y., Jiang, W., Zhang, Y., Li, X.-Y., 2024. EarSleep: In-ear acoustic-based physical and physiological activity recognition for sleep stage detection. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 8 (2).

Hassan, A.R., Subasi, A., 2017. A decision support system for automated identification of sleep stages from single-channel EEG signals. Knowl.-Based Syst. 128, 115–124.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Huang, C.-S., Lin, C.-L., Ko, L.-W., Liu, S.-Y., Su, T.-P., Lin, C.-T., 2014. Knowledge-based identification of sleep stages based on two forehead electroencephalogram channels. Front. Neurosci. 8, 263.

Innocenti, S.U., Becattini, F., Pernici, F., Del Bimbo, A., 2021. Temporal binary representation for event-based action recognition. In: 2020 25th International Conference on Pattern Recognition. ICPR, IEEE.

Kay, M., Choe, E.K., Shepherd, J., Greenstein, B., Watson, N., Consolvo, S., Kientz, J.A., 2012. Lullaby: a capture & access system for understanding the sleep environment. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. pp. 226–234.

Kemp, B., Zwinderman, A.H., Tuk, B., Kamphuisen, H.A., Oberye, J.J., 2000. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. IEEE Trans. Biomed. Eng. 47 (9), 1185–1194.

Khalighi, S., Sousa, T., Santos, J.M., Nunes, U., 2016. ISRUC-sleep: A comprehensive public dataset for sleep researchers. Comput. Methods Programs Biomed. 124, 180–192.

Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.B., 2016. Hots: a hierarchy of event-based time-surfaces for pattern recognition. IEEE Trans. Pattern Anal. Mach. Intell..

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems. pp. 6403–6414.

Lee, T., Cho, Y., Cha, K.S., Jung, J., Cho, J., Kim, H., Kim, D., Hong, J., Lee, D., Keum, M., et al., 2023. Accuracy of 11 wearable, nearable, and airable consumer sleep trackers: prospective multicenter validation study. JMIR MHealth UHealth 11 (1), e50983.

Liu, C., Qi, X., Lam, E.Y., Wong, N., 2022. Fast classification and action recognition with event-based imaging. IEEE Access 10, 55638–55649.

Liu, Q., Xing, D., Tang, H., Ma, D., Pan, G., 2021. Event-based action recognition using motion information and spiking neural networks.. In: IJCAI. pp. 1743–1749.

Liu, C., Xiong, J., Cai, L., Feng, L., Chen, X., Fang, D., 2019. Beyond respiration: Contactless sleep sound-activity recognition using RF signals. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3 (3).

López-Larraz, E., Escolano, C., Robledo-Menéndez, A., Morlas, L., Alda, A., Minguez, J., 2023. A garment that measures brain activity: proof of concept of an eeg sensor layer fully implemented with smart textiles. Front. Hum. Neurosci. 17, 1135153.

Matar, G., Lina, J.-M., Kaddoum, G., 2019. Artificial neural network for in-bed posture classification using bed-sheet pressure sensors. IEEE J. Biomed. Health Inform. 24 (1), 101–110.

Miao, S., Chen, G., Ning, X., Zi, Y., Ren, K., Bing, Z., Knoll, A., 2019. Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. Front. Neurorobotics 13, 38.

Mohammadi, S.M., Enshaeifar, S., Hilton, A., Dijk, D.-J., Wells, K., 2020. Transfer learning for clinical sleep pose detection using a single 2d IR camera. IEEE Trans. Neural Syst. Rehabil. Eng. 29, 290–299.

Moyen, N.E., Ediger, T.R., Taylor, K.M., Hancock, E.G., Holden, L.D., Tracy, E.E., Kay, P.H., Irick, C.R., Kotzen, K.J., He, D.D., 2024. Sleeping for one week on a temperature-controlled mattress cover improves sleep and cardiovascular recovery. Bioengineering 11 (4), 352.

Mueggler, E., Forster, C., Baumli, N., Gallego, G., Scaramuzza, D., 2015. Lifetime estimation of events from dynamic vision sensors. In: 2015 IEEE International Conference on Robotics and Automation. ICRA, IEEE.

Nakajima, K., Matsumoto, Y., Tamura, T., 2000. A monitor for posture changes and respiration in bed using real time image sequence analysis. In: Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No. 00CH37143), vol. 1, IEEE, pp. 51–54.

Neumann, L., Zisserman, A., Vedaldi, A., 2018. Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection. In: NeurIPS 2018 Workshop MLITS.

Parameshwara, C.M., Li, S., Fermüller, C., Sanket, N.J., Evanusa, M.S., Aloimonos, Y., 2021. Spikems: Deep spiking neural network for motion segmentation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 3414–3420.

Peng, Y., Zhang, Y., Xiong, Z., Sun, X., Wu, F., 2023. GET: Group event transformer for event-based vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6038–6048.

Piriyajitakonkij, M., Warin, P., Lakhan, P., Leelaarporn, P., Kumchaiseemak, N., Suwajanakorn, S., Pianpanit, T., Niparnan, N., Mukhopadhyay, S.C., Wilaiprasitporn, T., 2021. SleepPoseNet: Multi-view learning for sleep postural transition recognition using UWB. IEEE J. Biomed. Health Inform. 25 (4), 1305–1314.

Plou, C., Gallego, N., Sabater, A., Montijano, E., Urcola, P., Montesano, L., Martinez-Cantin, R., Murillo, A.C., 2024. EventSleep: Sleep activity recognition with event cameras. Wokshop Neuromorphic Vis. - Eur. Conf. Comput. Vis. Work..

Pradhan, B.R., Bethi, Y., Narayanan, S., Chakraborty, A., Thakur, C.S., 2019. N-HAR: A neuromorphic event-based human activity recognition system using memory surfaces. In: 2019 IEEE International Symposium on Circuits and Systems. ISCAS, IEEE, pp. 1–5.

Sabater, A., Montesano, L., Murillo, A.C., 2022. Event transformer. a sparse-aware solution for efficient event data processing. In: Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops.

Sabater, A., Montesano, L., Murillo, A.C., 2023. Event transformer+. a multi-purpose solution for efficient event data processing. IEEE Trans. Pattern Anal. Mach. Intell..

Sathyanarayana, S., Satzoda, R.K., Sathyanarayana, S., Thambipillai, S., 2018. Vision-based patient monitoring: a comprehensive review of algorithms and technologies. J. Ambient. Intell. Humaniz. Comput. 9, 225–251.

Shrestha, S.B., Orchard, G., 2018. SLAYER: Spike layer error reassignment in time. In: NeurIPS.

Sun, J., Zhang, Q., Wang, J., Cao, J., Cheng, H., Xu, R., 2025. Event masked autoencoder: Point-wise action recognition with event-based cameras. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5.

Terzano, M.G., Parrino, L., Sherieri, A., Chervin, R., Chokroverty, S., Guilleminault, C., Hirshkowitz, M., Mahowald, M., Moldofsky, H., Rosa, A., et al., 2001. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. Sleep Med. 2 (6), 537–554.

Vanderlinden, J., Boen, F., Van Uffelen, J., 2020. Effects of physical activity programs on sleep outcomes in older adults: a systematic review. Int. J. Behav. Nutr. Phys. Act. 17 (1), 1–15.

Vasudevan, A., Negri, P., Di Ielsi, C., Linares-Barranco, B., Serrano-Gotarredona, T., 2021. SL-animals-DVS: event-driven sign language animals dataset. Pattern Anal. Appl..

Vicente-Sola, A., Manna, D.L., Kirkland, P., Caterina, G.D., Bihl, T.J., 2025. Spiking neural networks for event-based action recognition: A new task to understand their advantage. Neurocomputing 611, 128657.

Wang, Z., Hu, Y., Liu, S.-C., 2022. Exploiting spatial sparsity for event cameras with visual transformers. In: 2022 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 411–415.

Wang, X., Wu, Z., Jiang, B., Bao, Z., Zhu, L., Li, G., Wang, Y., Tian, Y., 2024a. HARDVS: Revisiting human activity recognition with dynamic vision sensors. Proc. AAAI Conf. Artif. Intell. 38 (6), 5615–5623. http://dx.doi.org/10.1609/aaai.v38i6.28372, URL https://ojs.aaai.org/index.php/AAAI/article/view/28372.

Wang, Q., Xu, Z., Lin, Y., Ye, J., Li, H., Zhu, G., Ali Shah, S.A., Bennamoun, M., Zhang, L., 2024b. Dailydvs-200: A comprehensive benchmark dataset for event-based action recognition. In: European Conference on Computer Vision. Springer, pp. 55–72.

Wang, Q., Zhang, Y., Yuan, J., Lu, Y., 2019. Space-time event clouds for gesture recognition: From RGB cameras to event cameras. In: IEEE Winter Conf. on Applications of Computer Vision. WACV.

Weng, W., Zhang, Y., Xiong, Z., 2021. Event-based video reconstruction using transformer. In: Proc. of the IEEE/CVF Int. Conf. on Computer Vision.

Yadav, S.K., Tiwari, K., Pandey, H.M., Akbar, S.A., 2021. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. Knowl.-Based Syst. 223, 106970.

Zhang, J., Dong, B., Zhang, H., Ding, J., Heide, F., Yin, B., Yang, X., 2022. Spiking transformers for event-based single object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8801–8810.

Zhang, Y., Zhang, Z., Zhang, Y., Bao, J., Zhang, Y., Deng, H., 2019. Human activity recognition based on motion sensor using u-net. IEEE Access 7, 75213–75226.

Zhou, J., Zheng, X., Lyu, Y., Wang, L., 2024. Exact: Language-guided conceptual reasoning and uncertainty estimation for event-based action recognition and more. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18633–18643.