CGSSN: A Coupled Graph Spiking Skeleton Network for Human Action Recognition Using Neuromorphic Vision Sensors

Lihui Wu
School of Electrical and Electronic Engineering
The University of Sheffield
Sheffield, United Kingdom
lwu17@sheffield.ac.uk

Charith Abhayaratne
School of Electrical and Electronic Engineering
Centre for Machine Intelligence
The University of Sheffield
Sheffield, United Kingdom
c.abhayaratne@sheffield.ac.uk

***Abstract*—Event cameras provide sparse and asynchronous event streams with high temporal resolution, enabling human action recognition under fast motion and challenging illumination conditions. However, effective modeling of event streams for event-based human action recognition remains challenging due to polarity imbalance, timestamp jitter, and the inherent mismatch between structural reasoning and transient temporal dynamics. To address these challenges, we propose CGSSN, a coupled graph–spiking dual-path framework that jointly models structure-aware spatial relations and spike-driven temporal dynamics. CGSSN introduces a unified polarity-aware Poisson temporal encoding (UPPTE) module to enhance spike stability and robustness against temporal jitter, and adopts parallel graph and spiking pathways to capture complementary structural and transient cues. In addition, a lightweight cross-branch alignment and gated fusion strategy is incorporated to improve representation consistency across heterogeneous pathways. Experimental results on five public event-based human action recognition benchmarks demonstrate that CGSSN consistently outperforms existing CNN-based, Transformer-based, and spiking-based methods, achieving state-of-the-art performance on multiple datasets. Comprehensive Ablation studies further verify the effectiveness of each proposed components. The software package associated with this work is available on GitHub: https://github.com/lihuiwu97/CGSSN**

***Index Terms*—Event-based vision, human action recognition, spiking neural networks, graph neural networks, temporal encoding, neuromorphic computing.**

I. Introduction

Event cameras (neuromorphic vision sensors) asynchronously output brightness changes as event streams (x, y, t, p) [1] [2], where (x, y) denotes the pixel location, t is the timestamp, and $p \in \{+1, -1\}$ indicates the on/off polarity. Compared to conventional frame-based cameras, event streams are sparse and provide high temporal resolution, low latency, and a wide dynamic range, which makes them particularly suitable for human action recognition (HAR) under fast motion and challenging illumination conditions. Early event-based gesture recognition systems have demonstrated the feasibility of fully event-driven HAR pipelines operating directly on asynchronous events [3], [4]. However, irregular event sampling, polarity-related statistical variations, and the sensitivity of temporal encoding to noise and temporal misalignment continue to pose significant challenges for robust event-based HAR.

A key challenge lies in *polarity imbalance*. ON and OFF events encode opposite directions of intensity changes and capture complementary motion boundaries. However, polarity distributions are often highly skewed across different scenes and action categories, and can be further distorted by sensor bias or preprocessing operations. In the experiments, this imbalance often leads to dominant activations and unstable spiking behavior, which noticeably degrades generalization performance, particularly when event streams are converted into discrete representations or spike sequences.

Beyond polarity imbalance, *timestamp jitter and temporal uncertainty* also impact the stability of spike-based characterization. Event timestamps are susceptible to sensor noise and discretization strategies such as temporal binning or windowing. As a result, simple temporal encoding schemes, including direct Poisson or rate-based sampling, often fail to robustly represent event streams, since minor temporal perturbations can lead to significant variations in spike firing patterns. This sensitivity is particularly detrimental to HAR, where accurate discrimination between similar actions depends on fine-grained temporal cues.

A further challenge stems from *the inherent conflict between structural reasoning and transient temporal dynamics*. Many human actions exhibit clear structural patterns that can benefit from topology-aware modeling, such as skeleton or graph-based dependencies, while simultaneously containing transient activation bursts and fine-grained temporal cues that are difficult to preserve through graph propagation alone. In practice, graph-centering models often emphasize spatial aggregation and long-range coordination while neglecting micro-level temporal dynamics, whereas pure time-series spiking neural networks often fail to fully exploit the underlying spatial structure.

To address these challenges, we propose **CGSSN** (**C**oupled **G**raph–**S**piking **S**keleton **N**etwork), a spiking–graph hybrid framework that integrates robust polarity-aware temporal encoding with a coupled dual-path architecture. CGSSN empolys

a unified polarity-aware Poisson temporal encoding (**UPPTE**) to stabilize spiking representations under polarity imbalance and timestamp jitter. On top of this encoder, a *Graph Path* performs structure-aware spiking message passing, while a *Pure-SNN Path* focuses on modeling transient temporal dynamics. A lightweight cross-branch alignment mechanism is further introduced to encourage consistent representations prior to gated fusion.

Contributions. We summarize three contributions:

- **UPPTE: unified robust spiking encoding.** We introduce a modular polarity-aware encoder that improves robustness to timestamp jitter through robust-kernel smoothing prior to sampling, and stabilizes spiking activity via spike-rate normalization.
- **Coupled dual-path modeling.** We propose a parallel spiking–graph architecture that jointly captures topology-aware dependencies (Graph Path) and transient temporal dynamics (Pure-SNN Path).
- **Alignment and stable fusion.** We incorporate a lightweight cross-branch alignment (TW-Align) and gated fusion to reduce branch interference and improve representation consistency. [5]

The remainder of this paper is organized as follows. Section II reviews related work on event-based action recognition and spiking neural networks. Section III introduces the proposed CGSSN framework and details the coupled dual-path architecture as well as the unified polarity-aware temporal encoding. Section IV presents experimental settings and quantitative comparisons with state-of-the-art methods, followed by ablation studies. Finally, Section V concludes the paper and discusses future research directions.

II. Related Work

Early event-based human action recognition methods commonly convert asynchronous event streams into frame-like representations, such as event frames, voxel grids, or time surfaces, enabling the reuse of conventional neural network(CNN) or temporal convolution backbones [6], [7]. Although CNN-based approaches can use mature video recognition pipelines [8], their performance is highly sensitive to discretization hyperparameters, including temporal window size, bin number, and accumulation strategy. Inappropriate discretization may oversmooth fine-grained temporal cues or amplify background noise, resulting in unstable generalization across datasets and motion patterns.

Spiking neural networks (SNN) naturally align with event-based sensing, as both operate on impulse-driven temporal dynamics [9], [10]. Recent advances in surrogate-gradient learning have enabled large-scale spiking architectures for vision tasks, and several spiking-based HAR systems directly process event streams to preserve asynchronous timing information [3]. Nevertheless, temporal encoding remains a critical bottleneck. Conventional rate-based or Poisson sampling strategies, although computationally efficient, are highly sensitive to timestamp jitter, polarity distribution shifts, and discretization noise [11]. Even minor temporal perturbations can induce significant variations in spike firing patterns, leading to unstable discharge behavior and optimization difficulties.

Several studies stabilize spike statistics via normalization or learnable encoding [12], most process ON and OFF channels independently without unified normalization, limiting robustness under polarity imbalance and sensor bias.

Graph neural networks have been widely adopted in skeleton-based action recognition by propagating information along joint topology and capturing long-range spatial dependencies [13], [14]. Similarly, shape analysis on point cloud data has been inspired by modelling them on graph spectral domains [15]. However, in event-based settings, graph-centric propagation tends to emphasize spatial aggregation at the expense of short-lived temporal activations, which suppress fine-grained transient temporal cues that are critical for distinguishing visually similar actions.

Recent hybrid architectures attempt to combine spiking computation with graph reasoning or multi-stream designs to exploit complementary spatial and temporal representations [16]. State-space models further provide efficient long-range temporal modeling but often lack explicit spatial structure reasoning [17]. Despite their effectiveness, most existing hybrid approaches do not explicitly address polarity imbalance and timestamp jitter at the encoding stage, nor do they enforce representation consistency between structural and temporal branches,which can lead to cross-branch interference and unstable fusion.

III. The Proposed Method

A. Overview

CGSSN (Coupled Graph–Spiking Skeleton Network) is a spiking-graph hybrid framework for event-based human action recognition (HAR). The framework is guided by two key principles: *robust spiking encoding* to mitigate polarity imbalance and timestamp jitter, and *coupled structural–dynamic modeling* to jointly capture topology-aware dependencies and transient temporal cues.

As illustrated in Fig. 1, CGSSN is organized into three sequential stages:

1) **UPPTE encoder**: converts event-derived polarity features into stable spiking features with enhanced robustness to polarity imbalance and temporal jitter.
2) **Coupled dual-path backbone**: a **Graph Path** performs structure-aware spiking message passing over an event-derived topology, while a **Pure-SNN Path** focuses on modeling transient temporal dynamics through spiking computation.
3) **Alignment & fusion**: a lightweight cross-branch alignment mechanism reduces representation mismatch between the two pathways, followed by gated fusion for final classification.

B. Initial Event Representation

Given an event stream $\mathcal{E} = \{(x_i, y_i, t_i, p_i)\}_{i=1}^{M}$,where $p_i \in \{+1, -1\}$ denotes the event polarity, we discretize time axis into T bins [18]. We further construct N nodes

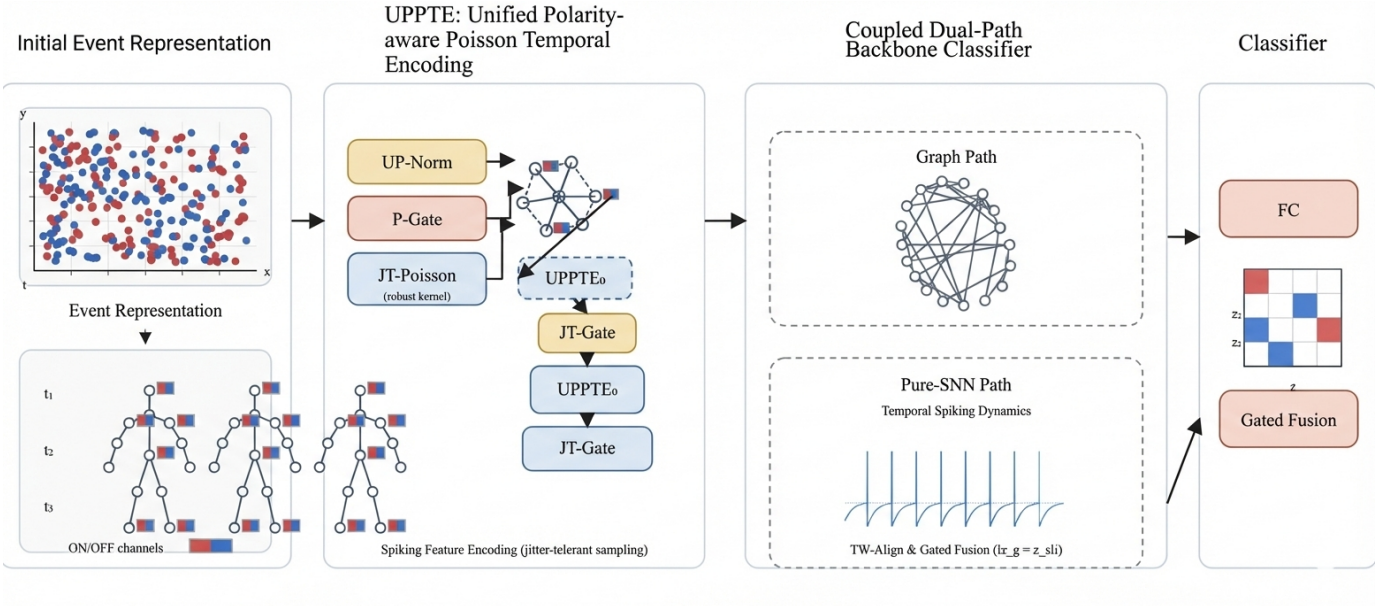

Fig. 1. Overall framework of CGSSN. The network first constructs an initial event representation from asynchronous event streams and applies unified polarity-aware Poisson temporal encoding (UPPTE) to obtain stable spiking features. These features are processed by a coupled dual-path backbone consisting of a Graph Path for structure-aware modeling and a Pure-SNN Path for transient temporal dynamics. Cross-branch alignment and gated fusion are then applied before final classification.

per sequence, corresponding to predefined structural elements such as skeleton joints or spatial regions. For each time bin $t \in \{1, \ldots, T\}$ and node $n \in \{1, \ldots, N\}$, ON and OFF events are aggregated separately to form a spatiotemporal feature tensor:

$$\begin{aligned} X &\in \mathbb{R}^{T\times N\times 2}, \\ X(t,n,0) &= \text{ON-agg}(t,n), \\ X(t,n,1) &= \text{OFF-agg}(t,n). \end{aligned} \tag{1}$$

where T denotes the number of temporal bins, N denotes the number of graph nodes, and the last dimension corresponds to ON/OFF polarity channels.

This representation explicitly separates ON/OFF channels to preserve polarity semantics, which is critical for capturing event-based motion boundaries. Moreover, the (T, N) organization provides a unified interface for both branches: the Graph Path expects node-wise features at each time step, whereas the Pure-SNN Path processes a temporally ordered feature sequence. In practice, ON-agg and OFF-agg operators can be instantiated using simple event counts or normalized accumulation schemes, such as polarity-weighted histograms. CGSSN does not rely on a specific aggregation strategy; this stage is intentionally kept lightweight to remain compatible with avrious event representations, while robustness to polarity imbalance and temporal noise is explicitly delegated to the subsequent UPPTE module.

C. UPPTE: Unified Polarity-aware Poisson Temporal Encoding

In initial experiments, we observed that even minor temporal perturbations frequently induce unstable spikes, prompting us to develop anti-jitter encoding mechanisms. *UPPTE* is designed to produce stable and robust spiking representations from event-derived polarity features by explicitly addressing polarity imbalance, timestamp jitter, and firing instability during temporal encoding.To clarify the internal processing flow of the proposed unified polarity-aware Poisson temporal encoding, Fig. 2 illustrates the sequential operations within the UPPTE module. Given the input tensor X, UPPTE maps it into spiking features $H \in \mathbb{R}^{T\times N\times D}$, where D denotes the spiking feature embedding dimension.

1) UP-Norm: Unified Polarity Normalization: ON and OFF polarity channels often exhibit different scales and statistical distributions across scenes and actions, for example, due to illumination variations and background dynamics. If one polarity dominates in magnitude, the resulting imbalance can bias spiking activations and degrade robustness under domain shifts. We first normalize ON/OFF values to a shared amplitude domain:

$$\tilde{X}(t,n,c) = \frac{X(t,n,c)}{\max_{t,n,c} X(t,n,c) + \epsilon}, \quad c \in \{0,1\}, \tag{2}$$

where c denotes the polarity channel index(c = 0 and c = 1) corresponding to OFF and ON event channels, respectively, and ϵ is a small constant for numerical stability.

2) P-Gate: Learnable Polarity Gate: **Motivation.** Even after normalization, the relative *importance* of ON and OFF polarity can vary across datasets and action categories, making fixed fusion rules suboptimal.

Operation. We introduce a learnable gate to adaptively balance ON/OFF contributions:

$$\begin{aligned} U(t,n) &= g\,\tilde{X}(t,n,0) + (1-g)\,\tilde{X}(t,n,1), \\ &\text{where } g = \sigma(\theta_g) \in (0,1). \end{aligned} \tag{3}$$

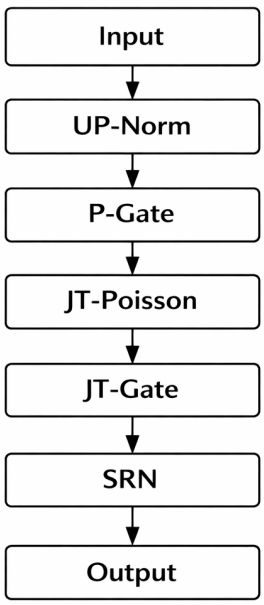

Fig. 2. Internal processing flow of the unified polarity-aware Poisson temporal encoding (UPPTE). The input event-derived features are sequentially processed by polarity normalization (UP-Norm), learnable polarity gating (P-Gate), jitter-tolerant Poisson sampling (JT-Poisson), jitter-tolerant temporal gating (JT-Gate), and spike-rate normalization (SRN) to produce stable spiking representations.

$U(t,n)$ denotes the unified polarity-aware activation at time bin t and node n, $\sigma(\cdot)$ is the sigmoid function, and θ_g is a learnable scalar parameter. P-Gate models polarity preference and prevents the network from being forced into a fixed polarity mixing regime.

3) JT-Poisson: Jitter-Tolerant Poisson Sampling with Robust Temporal Kernel: Directly sampling spikes from instantaneous activations is highly sensitive to timestamp jitter, as minor temporal shifts can significantly alter spike firing patterns. To mitigate this issue, we smooth the activation signals prior to spike sampling by aggregating evidence from neighboring time bins using a robust temporal kernel. This operation produces a jitter-tolerant firing intensity, which is then used to generate stochastic spikes. Specifically, we compute the firing intensity as

$$\lambda(t,n) = \sum_{\tau=1}^{T} U(\tau,n)\,\kappa_{\text{rob}}(t-\tau), \tag{4}$$

where $\lambda(t,n)$ represents the expected firing intensity of node n at time bin t. κ_{rob} is a normalized non-negative temporal kernel satisfying $\sum_\tau \kappa_{\text{rob}}(t-\tau) = 1$. In this work, we instantiate κ_{rob} as a Laplacian kernel:

$$\kappa_{rob}(\Delta\tau) \propto \exp\left(-\frac{|\Delta\tau|}{b}\right), \tag{5}$$

where b controls the temporal support of the kernel. Larger values of b yield wider temporal tolerance and stronger smoothing across neighboring bins.

Given the firing intensity $\lambda(t,n)$, we generate stochastic spikes using a Bernoulli approximation of an inhomogeneous Poisson process:

$$S(t,n) \sim \text{Bernoulli}(1-\exp(-\lambda(t,n)\delta t)), \tag{6}$$

where $S(t,n) \in \{0,1\}$ denotes the sampled spike of node n at time bin t, and δt denotes the temporal discretization step.

When an informative event cluster undergoes a small temporal shift, the resulting firing intensity $\lambda(t,n)$ remains stable because the robust kernel redistributes activation mass across neighboring bins. As a result, spike generation becomes less sensitive to minor temporal perturbations, effectively reducing the impact of timestamp jitter and time-warp effects. This temporal smoothing mechanism improves the stability and robustness of spiking representations.

4) JT-Gate: Jitter-Tolerant Temporal Gating: Although *JT-Poisson* produces a smoothed firing intensity that is robust to timestamp jitter, the resulting spike sequences may still contain spurious activations in temporally ambiguous regions where the firing intensity $\lambda(t,n)$ is close to the decision boundary. To selectively suppress unreliable spike outputs and improve the signal-to-noise ratio of the spiking representation, we introduce a lightweight temporal gating mechanism applied after each Poisson sampling stage.

Specifically, JT-Gate computes a scalar gate value for each node n at time bin t based on the local firing intensity:

$$g(t,n) = \sigma(W_g\,\lambda(t,n) + b_g), \tag{7}$$

where W_g and b_g are learnable scalar parameters, and $\sigma(\cdot)$ denotes the sigmoid activation function. The gated spike representation is then obtained as:

$$\tilde{S}(t,n) = g(t,n) \odot S(t,n), \tag{8}$$

where $S(t,n) \in \{0,1\}$ is the Bernoulli-sampled spike, and $\odot$ denotes element-wise multiplication. JT-Gate effectively acts as a confidence filter: nodes with high and stable firing intensity (large λ) receive gate values close to 1 and pass through unmodified, while nodes in temporally uncertain regions are adaptively down-weighted.

This mechanism is applied twice within UPPTE — once after the initial JT-Poisson sampling stage (yielding UPPTE_0) and once after the second encoding stage — as illustrated in Fig. 2. The parameters W_g and b_g are shared across both applications to keep the module lightweight and parameter-efficient.

5) SRN: Spike-Rate Normalization for Firing Stability: To prevent pathological firing behavior and enforce balanced spiking activity, we regulate node-wise firing statistics using spike-rate normalization. We compute the node-wise average firing rate as

$$r(n) = \frac{1}{T}\sum_{t=1}^{T} S(t,n), \tag{9}$$

which measures the average firing rate (spike probability) of node n over the temporal window.

In this work, SRN is applied as a lightweight normalization mechanism to regulate spike statistics. Specifically, we use $r(n)$ as a stability reference to prevent pathological firing patterns and encourage consistent activation distributions across nodes.

6) Spiking Feature Construction: The final spiking feature is constructed by modulating learnable embeddings with sampled spike masks:

$$F(t,n) = \phi([\tilde{X}(t,n,0), \tilde{X}(t,n,1)]) \in \mathbb{R}^D, \quad (10)$$

$$H(t,n) = S(t,n) \odot F(t,n), \quad (11)$$

where $\phi(\cdot)$ denotes a linear layer, and D is the feature dimension. $[\cdot,\cdot]$ denotes channel-wise concatenation of OFF and ON polarity features. $H(t,n)$ denotes the spike-modulated feature representation at time t and node n. The element-wise product ($\odot$) enforces spike-driven temporal sparsity while preserving polarity-aware feature representations, enabling efficient spiking computation and structured temporal modeling.

D. Coupled Dual-Path Backbone

After UPPTE encoding, CGSSN processes the spatio-temporal spiking representation $\mathbf{H} \in \mathbb{R}^{T\times N\times D}$ through two complementary pathways: a *Graph Path* for structure-aware spiking message passing and a *Pure-SNN Path*. for modeling transient temporal dynamics. [19] The two pathways operate in parallel and are explicitly coupled to produce a unified action representation.

1) Graph Path: Structure-aware Spiking Message Passing: We construct a node graph $G = (V, E)$ with $|V| = N$. The adjacency matrix is defined using a fixed skeleton topology and remains unchanged during training and inference. The adjacency matrix $A \in \mathbb{R}^{N\times N}$ encodes the connectivity between nodes and remains unchanged across training and inference to avoid introducing additional temporal instability in spiking propagation.

At each time step t, node features are updated by aggregating neighborhood information using an attention-based formulation adapted to spiking representations:

$$\alpha_{ij}(t) = \mathrm{softmax}_{j\in\mathcal{N}(i)}\left(\frac{q_i(t)^\top k_j(t)}{\sqrt{D}}\right), \quad (12)$$

$$h'_i(t) = \sum_{j\in\mathcal{N}(i)} \alpha_{ij}(t)\, v_j(t), \quad (13)$$

where $q_i(t)$, $k_j(t)$, and $v_j(t)$ denote linear projections of the spiking node features $H(t,\cdot)$, and $\mathcal{N}(i)$ denotes the neighborhood set of node i, the softmax normalization is performed over $\mathcal{N}(i)$, D denotes the spiking feature embedding dimension, and $h'_i(t)$ represents the updated spiking node feature after message aggregation. This formulation enables content-adaptive propagation while preserving spike-driven sparsity.

Temporal and spatial pooling are applied to obtain a global structure-aware embedding:

$$\mathbf{z}_g = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N} h'_i(t)\right) \in \mathbb{R}^D. \quad (14)$$

where z_g denotes the global graph-path embedding. The inner summation performs spatial pooling across nodes, while the outer summation aggregates temporal information.

2) Pure-SNN Path: Transient Dynamics Modeling: Node-wise spiking features are first aggregated to form a compact temporal sequence:

$$\bar{\mathbf{h}}(t) = \frac{1}{N}\sum_{n=1}^{N}\mathbf{H}(t,n) \in \mathbb{R}^D. \quad (15)$$

where $\bar{h}(t)$ denotes the temporally ordered node-averaged spiking feature representation at time bin t. N denotes the number of graph nodes. The resulting sequence is processed by a temporal spiking stack composed of Leaky Integate-and-fire(LIF) based layers. A standard LIF neuron update is given by:

$$u(t) = \beta u(t-1) + W\bar{\mathbf{h}}(t) - v_{th}s(t-1), \quad (16)$$

$$s(t) = \Theta(u(t) - v_{th}), \quad (17)$$

where $u(t)$ denotes the membrane potential, β is the leak factor, W is a learnable weight matrix, v_{th} denotes the firing threshold, $s(t)$ is the output spike, and $\Theta(\cdot)$ denotes the Heaviside step function with surrogate gradients during training. Temporal pooling over time yields a dynamics-aware embedding:

$$\mathbf{z}_s = \frac{1}{T}\sum_{t=1}^{T} s(t) \in \mathbb{R}^D. \quad (18)$$

where z_s denotes the global embedding produced by the Pure-SNN Path.

3) Cross-Path Coupling: The two pathways are explicitly coupled through a temporal alignment and gated fusion mechanism (described in the following section). This coupling enforces consistency between structure-aware and dynamics-aware representations while preserving their complementary properties. As a result, CGSSN jointly models coordinated spatial patterns and transient temporal dynamics in a unified spiking framework.

Fig. 3 provides an abstract illustration of the coupled dual-path backbone, highlighting the parallel processing of structural and temporal spiking representations and their subsequent alignment and fusion.

E. TW-Align and Gated Fusion

Directly fusing such asynchronous embeddings can lead to phase mismatch and scale inconsistency, thereby degrading fusion effectiveness and destabilizing training.

Temporal-wise alignment(TW-Align). To address this discrepancy, we introduce a lightweight temporal-wise alignment objective that encourages consistent global representations while preserving branch complementarity:

$$\mathcal{L}_{\text{align}} = \|\mathbf{z}_g - \mathbf{z}_s\|_1, \quad (19)$$

where z_g and z_s denote the structure-aware and dynamics-aware embeddings of the same sample, and $|\cdot|_1$ denotes the ℓ_1 norm.

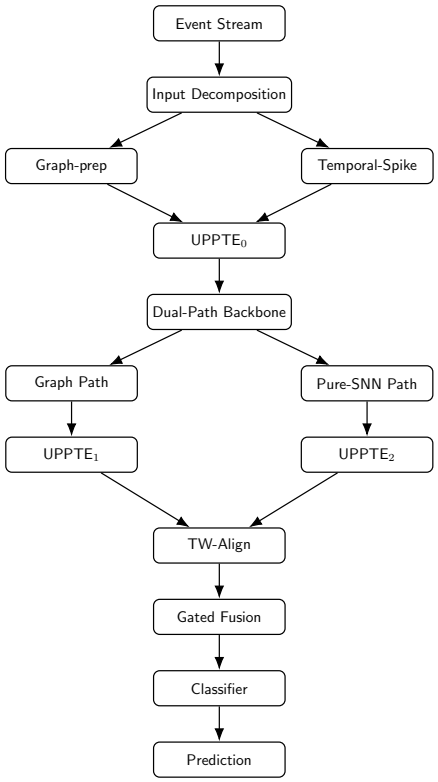

Fig. 3. Abstract illustration of the coupled dual-path backbone in CGSSN. After polarity-aware temporal encoding, spiking features are processed in parallel by a Graph Path for structure-aware message passing and a Pure-SNN Path for temporal spiking dynamics. The two pathways are aligned and fused to form a unified representation for action classification.

This alignment loss serves as a soft regularization rather than a hard constraint. It stabilizes the scale and phase of cross-branch representations without forcing the two pathways to collapse into identical features, thereby maintaining their complementary roles.

Gated fusion. We employ a learnable gating mechanism to adaptively balance the contributions of the two branches on a per-sample basis:

$$\alpha = \sigma\left(W_f[\mathbf{z}_g; \mathbf{z}_s]\right), \tag{20}$$

$$\mathbf{z} = \alpha \odot \mathbf{z}_g + (1-\alpha) \odot \mathbf{z}_s, \tag{21}$$

where $[\cdot;\cdot]$ denotes concatenation, $\odot$ represents element-wise multiplication, $\sigma(\cdot)$ is the sigmoid activation function, W_f is a learnable gating projection matrix, and α is a scalar gating coefficient. The gating mechanism provides flexible, sample-adaptive fusion, allowing the model to dynamically emphasizee structure-aware or dynamics-aware representations depending on the input characteristics.

Classifier. The fused representation is finally fed into a linear classifier:

$$\hat{y} = \mathrm{softmax}(W_c \mathbf{z}), \tag{22}$$

to produce the action prediction, where W_c is the classifier weight matrix.

TABLE I
DATASET SUMMARY USED IN OUR EXPERIMENTS.

Dataset	#Classes	Scale	Task
DailyDVS-200 [20]	200	22,000	Action
DvsGesture [3]	11	29,000	Action
DailyAction [21]	12	1,200	Action
THU E-ACT-50-CHL [22]	50	2,330	Action
CeleX-HAR [23]	150	124,625	Action

F. Training Objective and Notation

The network is trained end-to-end using a weighted composite loss function:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{rate}}\mathcal{L}_{\text{rate}}, \tag{23}$$

where $\mathcal{L}_{cls}$ denotes the standard cross-entropy classification loss, $\mathcal{L}_{\text{align}}$ enforces representation consistency between the two branches as defined in Eq. (17), and $\mathcal{L}_{\text{rate}}$ regularizes spike statistics to prevent pathological firing behavior. The coefficients $\lambda align$ and λ_{rate} control the relative contributions of the auxiliary loss terms.

All hyperparameters, such as temporal bin width Δt, kernel decay factor b, and feature dimension D, are fixed across datasets and detailed in Section IV-B (Implementation Details).

IV. Performance Evaluation

A. Dataset and Evaluation Metric

We evaluate CGSSN on five event-based human action recognition benchmarks: **DailyDVS-200** [20], **DvsGesture** [3], **DailyAction** [21], **THU E-ACT-50-CHL** [22], and **CeleX-HAR** [23]. A summary of these datasets is provided in Table I. We report **Top-1 accuracy** (%) on the official test split for evaluation.

B. Implementation Details

We implement CGSSN using PyTorch and train the network end-to-end with surrogate-gradient optimization. Unless otherwise specified, all hyperparameters are shared across datasets.

a) Temporal discretization and encoding.: We discretize each event sequence into $T = 30$ bins with bin width $\Delta t =$ 10ms. In UPPTE, we use the jitter-tolerant temporal kernel κ_{rob} instantiated as a Laplacian kernel with decay parameter $b = 20$ms.

b) Graph construction and node definition.: We use $N = 25$ nodes following the standard 25-joint human skeleton topology of NTU RGB+D [24], widely adopted in skeleton-based HAR. Raw event streams are mapped onto these nodes by spatially binning events into 25 regions centered on projected joint positions, with ON/OFF events aggregated per temporal bin to form $\mathbf{X} \in \mathbb{R}^{T\times N\times 2}$. The adjacency matrix encodes physical joint connectivity with neighborhood size 4.

c) Training setup.: We train for 200 epochs using Adam with initial learning rate 1×10^{-3} and a step decay schedule at epochs 120,160. The batch size is 32. Loss weights are set as $\lambda_{\text{align}} = 0.1$ and $\lambda_{\text{rate}} = 1 \times 10^{-3}$. All experiments are conducted on NVIDIA RTX 3090.

C. Comparison with SOTA

We compare CGSSN with representative state-of-the-art methods spanning CNN-based, Transformer-based, state-space models, and spiking neural networks on five widely used event-based action recognition benchmarks. The quantitative results are summarized in Table II.

a) DailyDVS-200: On the DailyDVS-200 dataset, CGSSN achieves an accuracy of 57.72%, significantly outperforming CNN-based TSM (40.87%), Transformer-based Swin-T (48.06%), SNN Transformer Spikformer (36.94%), and the state-space model EVMamba (49.65%). The performance gain demonstrates the advantage of jointly modeling graph-structured spatial relations and spiking temporal dynamics for large-scale daily action scenarios captured by event cameras.

b) THU E-ACT-50-CHL: For the multi-view THU E-ACT-50-CHL benchmark, CGSSN obtains 73.85% accuracy, surpassing previous multi-view event-based methods such as MVF-Net (56.50%) and MATRIX-HAR (65.80%). This improvement indicates that the proposed coupled dual-path architecture effectively captures view-consistent motion patterns while maintaining robustness to viewpoint variations.

c) CeleX-HAR: On the CeleX-HAR dataset, CGSSN reaches 82.35%, outperforming EVMamba (72.30%) by a large margin. This result highlights the benefit of incorporating spiking graph representations and polarity-aware temporal encoding when dealing with high-resolution and high-frequency event streams.

d) DvsGesture and DailyAction-DVS: CGSSN achieves 99.92% on DvsGesture and 99.70% on DailyAction-DVS, outperforming HATS (98.70% / 93.80%) and SpikeVision (99.30% / 95.30%), demonstrating strong generalization across gesture and daily action scenarios. These results collectively demonstrate that CGSSN generalizes robustly across diverse event-based benchmarks, motion scales, and dataset sizes.

D. Ablation Study

We conduct extensive ablation studies on DailyDVS-200 and CeleX-HAR to analyze the contribution of each key component in CGSSN. The results are summarized in Table III.

a) Effect of UPPTE: Replacing the proposed unified polarity-aware Poisson temporal encoding (UPPTE) with the baseline PPTE leads to a notable performance drop on both datasets. Specifically, the accuracy decreases from 57.72% to 55.40% on DailyDVS-200 and from 82.35% to 80.20% on CeleX-HAR. This confirms that the unified polarity normalization and robust temporal sampling in UPPTE play a critical role in stabilizing spike representations and enhancing temporal discriminability.

b) Impact of jitter-tolerant Poisson sampling: Removing the JT-Poisson module results in consistent degradation, yielding 56.32% on DailyDVS-200 and 81.05% on CeleX-HAR. The performance drop indicates that the proposed jitter-tolerant temporal kernel effectively mitigates timestamp noise and event irregularity, leading to more robust spiking feature generation.

c) Effect of SRN and polarity modules: Removing SRN causes moderate drops to 57.15% and 81.72%, confirming its role in stabilizing firing distributions. Removing both P-Gate and UP-Norm decreases accuracy to 56.97% and 81.40%, highlighting the importance of unified polarity normalization for preserving motion contrast under ON/OFF imbalance.

d) Contribution of the pure-SNN path: Discarding the pure-SNN temporal dynamics branch leads to a noticeable performance drop to 56.55% on DailyDVS-200 and 81.10% on CeleX-HAR. This confirms that the temporal spiking dynamics branch provides complementary information beyond graph-based spatial modeling, and that the coupled dual-path design is necessary to fully exploit the spatiotemporal characteristics of event streams.

e) Effect of cross-branch alignment: When the temporal-wise alignment (TW-Align) module is removed, the accuracy decreases to 56.92% on DailyDVS-200 and 81.55% on CeleX-HAR. This indicates that enforcing representation consistency between the graph path and the pure-SNN path improves feature coherence and fusion effectiveness, leading to better final recognition performance.

f) Overall analysis: Overall, the full CGSSN configuration achieves the best performance on both datasets. Each component contributes positively to the final accuracy, while the largest performance drops are observed when removing UPPTE and the pure-SNN branch, highlighting the central role of robust temporal encoding and dual-path spiking modeling in the proposed framework.

V. Conclusions

We have proposed CGSSN, a coupled graph–spiking framework for event-based HAR that integrates UPPTE for polarity-aware, jitter-robust spike encoding with a dual-path backbone and cross-branch alignment and gated fusion. Experiments on five benchmarks demonstrate consistent state-of-the-art performance and ablation studies verify each component's contribution. Future work will explore multi-modal extensions and hardware-efficient implementations on neuromorphic platforms.

References

[1] 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*, vol. 44, no. 1, pp. 154–180, 2020.

[2] C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco, and T. Delbruck, "Retinomorphic event-based vision sensors: bioinspired cameras with spiking output," *Proceedings of the IEEE*, vol. 102, no. 10, pp. 1470–1484, 2014.

[3] 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*, 2017, pp. 7243–7252.

[4] S. Al-Obaidi, H. Al-Khafaji, and C. Abhayaratne, "Making sense of neuromorphic event data for human action recognition," *IEEE Access*, vol. 9, pp. 82 686–82 700, 2021.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

TABLE II
TOP-1 ACCURACY (%) COMPARISON ON EVENT-BASED ACTION RECOGNITION BENCHMARKS (OFFICIAL SPLITS).

Method	Category	DailyDVS-200	THU E-ACT-50-CHL	CeleX-HAR	DvsGesture	DailyAction-DVS
TSM [8]	CNN-based	40.87	–	–	–	–
Swin-T [25]	Transformer	48.06	–	–	–	–
Spikformer [26]	SNN Transformer	36.94	–	–	–	–
EVMamba [23]	State Space Model	49.65	–	72.30	–	–
MVF-Net [27]	Multi-view CNN	–	56.50	–	–	–
MATRIX-HAR [28]	Adaptive Temporal Resolution	–	65.80	–	–	–
HATS [6]	Event Surface CNN	–	–	–	98.70	93.80
SpikeVision [29]	SNN Transformer	–	–	–	99.30	95.30
CGSSN (Ours)	Graph-SNN Dual-path	**57.72**	**73.85**	**82.35**	**99.92**	**99.70**

TABLE III
ABLATION RESULTS OF CGSSN (TOP-1 ACC., %).

Variant	DailyDVS-200	CeleX-HAR
CGSSN (Full)	57.72	82.35
UPPTE $\rightarrow$ PPTE	55.40	80.20
w/o JT-Poisson	56.32	81.05
w/o SRN	57.15	81.72
w/o P-Gate & UP-Norm	56.97	81.40
w/o Pure-SNN Path	56.55	81.10
w/o TW-Align	56.92	81.55

[6] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, "HATS: histograms of averaged time surfaces for robust event-based object classification," in *CVPR*, 2018.

[7] 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*, 2019, pp. 5633–5643.

[8] J. Lin, C. Gan, and S. Han, "TSM: temporal shift module for efficient video understanding," in *ICCV*, 2019.

[9] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 51–63, 2019.

[10] Y. Wu, L. Deng, G. Li, J. Zhu, Y. Xie, and L. Shi, "Direct training for spiking neural networks: Faster, larger, better," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 1311–1318.

[11] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers in computational neuroscience*, vol. 9, p. 99, 2015.

[12] W. Fang, Z. Yu, Y. Chen, T. Huang, T. Masquelier, and Y. Tian, "Deep residual learning in spiking neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 056–21 069, 2021.

[13] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[14] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 026–12 035.

[15] B. Alwaely and C. Abhayaratne, "GHOSM: graph-based hybrid outline and skeleton modelling for shape recognition," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2s, pp. 1–23, 2023.

[16] Z. Zhu, J. Peng, J. Li, L. Chen, Q. Yu, and S. Luo, "Spiking graph convolutional networks," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 2434–2440, main Track.

[17] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[18] H. Li and C. Abhayaratne, "AW-GATCN: adaptive weighted graph attention convolutional network for event camera data joint denoising and object recognition," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.

[19] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[20] Q. Wang, Z. Xu, Y. Lin, J. Ye, H. Li, G. Zhu, S. A. A. Shah, M. Bennamoun, and L. Zhang, "DailyDVS-200: a comprehensive benchmark dataset for event-based action recognition (supplementary material)."

[21] Q. Liu, D. Xing, H. Tang, D. Ma, and G. Pan, "Event-based action recognition using motion information and spiking neural networks." in *IJCAI*, 2021, pp. 1743–1749.

[22] Y. Gao, J. Lu, S. Li, N. Ma, S. Du, Y. Li, and Q. Dai, "Action recognition and benchmark using event cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14 081–14 097, 2023.

[23] X. Wang, S. Wang, P. Shao, L. Zhu, B. Jiang, and Y. Tian, "Event stream based human action recognition: a high-definition benchmark dataset and algorithms," *International Journal of Computer Vision*, vol. 134, no. 4, p. 181, 2026.

[24] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: a large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.

[26] Z. Zhou, Y. Zhu, C. He, Y. Wang, S. YAN, Y. Tian, and L. Yuan, "Spikformer: When spiking neural network meets transformer," in *The Eleventh International Conference on Learning Representations*.

[27] Y. Deng, H. Chen, and Y. Li, "MVF-Net: a multi-view fusion network for event-based object classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8275–8284, 2021.

[28] M. H. Zafar, S. K. R. Moosavi, and F. Sanfilippo, "MATRIX-HAR: Lightweight temporal motion-guided feature network with adaptive temporal resolution based transfer learning for resource-constrained event camera-based human action recognition," *Expert Systems with Applications*, vol. 303, p. 130628, 2026.

[29] Z. Shen and F. Corradi, "Spike vision: A fully spiking neural network transformer-inspired model for dynamic vision sensors," in *2024 58th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2024, pp. 1537–1541.