EventGait: Towards Robust Gait Recognition with Event Streams
Supplementary Material

A. Ablation on the Number of Slices K

We first study the effect of the number of slices K used in the event representation (Section 3.1). K determines the temporal granularity fed into the Dynamic Motion Stream.

As shown in Table 1, performance is suboptimal when K is too small (e.g., $K = 2$), as fine-grained temporal dynamics critical for gait are lost. When K is too large (e.g., $K = 8$), the per-bin event slices become excessively sparse, which can introduce noise and slightly degrade performance. We found that $K = 6$ provides the best trade-off between temporal fidelity and signal density, achieving the highest overall accuracy. Therefore, we use $K = 6$ for all our main experiments.

Table 1. Ablation on the number of temporal bins K for event representation. All experiments are conducted on SUSTech1K-E.

Temporal Bins K	SUSTech1K-E			
	NM	CL	NT	Overall
$K = 2$	87.0	70.9	79.8	87.9
$K = 4$	90.4	78.0	83.2	91.2
$K = 6$	**92.5**	**78.1**	**84.8**	**92.8**
$K = 8$	90.1	76.4	80.4	89.3

B. Ablation on MoSE Configuration

We conduct a comprehensive ablation study analyzing the impact of membrane time constants (τ) on recognition performance (Section 3.2 and Figure 3).

Single Expert vs. Mixture of Experts. In Tab. 2, we first evaluate single-expert baselines initialized with fixed time constants ranging from $\tau = 2$ to $\tau = \infty$ (where ∞ denotes a standard Integrate-and-Fire neuron without leakage). The results reveal two key observations: (1) No single τ configuration achieves optimal performance across all covariates. For instance, while $\tau = 5$ yields decent results, it still lags behind mixture models in complex scenarios like Night (NT). (2) All MoSE configurations significantly outperform the single-expert baselines. This empirically validates our hypothesis that an ensemble of neurons with diverse temporal sensitivities can synergistically capture multi-scale dynamic patterns, offering greater robustness against illumination and motion variations.

Table 2. **Ablation on MoSE expert configurations.** We compare single experts with fixed time constants τ against various Mixture of Spiking Experts (MoSE) combinations. The results demonstrate that diverse temporal experts are essential for robust recognition. The configuration $\tau = \{2, 3, 5\}$ is selected for the best trade-off between performance and efficiency.

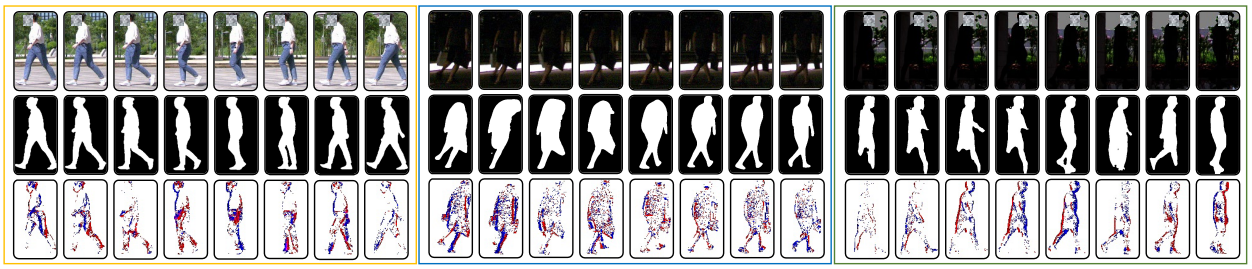
Model	**Expert Configuration (τ values)**	**SUSTech1K-E Rank-1 (%)**			
		NM	CL	NT	Overall
Single Expert	$\{2\}$	86.6	70.3	78.3	88.4
	$\{3\}$	88.2	72.5	79.7	88.6
	$\{5\}$	88.5	74.2	81.3	88.8
	$\{7\}$	87.7	75.6	82.3	87.6
	$\{\infty\}$ (IF Neuron)	88.3	73.9	81.5	87.5
MoSE (Ours)	$\{2, 5\}$	89.2	73.9	81.5	89.8
	$\mathbf{\{2, 3, 5\}}$	**92.5**	78.1	84.8	**92.8**
	$\{2, 3, 5, 7\}$	92.4	**78.9**	**85.1**	92.7

Optimal Configuration Selection. We further investigate different combinations of experts. As shown in the bottom rows of Tab. 2, the performance improves as the diversity of experts increases. The combination $\tau = \{2, 3, 5\}$ achieves a remarkable 92.8% overall rank-1 accuracy. Although adding a fourth expert ($\tau = \{2, 3, 5, 7\}$) provides a marginal gain in specific subsets (e.g., +0.3% in NT), the overall improvement saturates. To balance recognition accuracy with computational efficiency, we adopt $\{2, 3, 5\}$ as the default configuration for our final model.

C. Visualization of Feature Representations

We visualize the feature maps of the dual-stream architecture in Figure 2. We compare the raw RGB frames (Row a) and standard silhouettes (Row b) against the learned features from our Static Shape Stream (Row c) and Dynamic Motion Stream (Row d).

As shown in Row (c), the *Static Shape Stream* exhibits high activation across the entire human body region. Despite the input events being spatially sparse, the encoder (guided by our Cross-modal Structure Alignment) success-



(a) (b) (c)

Figure 1. **Visual comparison across modalities.** Rows from top to bottom: RGB, Silhouettes, and Events. (a) **Normal Light**. (b) **Night (NT)**. (c) **Synthesized Low-Light**. Unlike RGB and Silhouettes which degrade in low-light conditions, **Events** consistently retain robust gait cues.

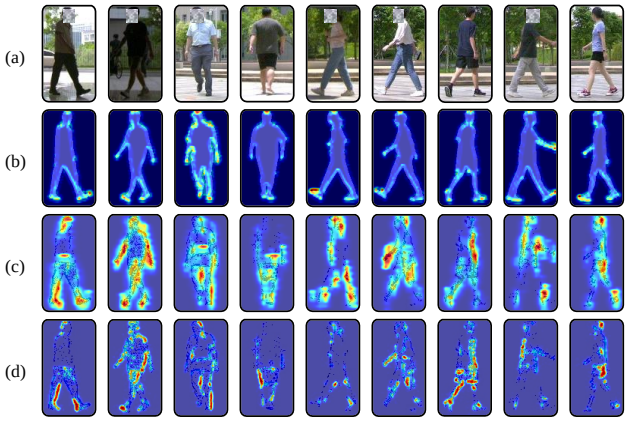


Figure 2. **Visualization of feature heatmaps.** (a) Raw RGB frames. (b) Feature heatmaps from silhouette methods [1]. (c) Feature heatmaps from our **Static Shape Stream**, which captures dense structural information. (d) Feature heatmaps from our **Dynamic Motion Stream**, which focuses on high-frequency motion cues like swinging limbs. The comparison highlights the complementary nature of our dual-stream design.

fully extracts key structural representations. In contrast, the *Dynamic Motion Stream* in Row (d) demonstrates a distinct focus on motion-salient areas. The heatmaps specifically highlight the high-frequency moving parts, such as swinging limbs and body contours, capturing the gait pattern.

D. Event Synthesis Pipeline

As illustrated in Figure 3, our synthesis pipeline converts RGB videos into high-fidelity event streams through a multi-stage process. First, to ensure motion continuity, we employ **Frame Interpolation** to upsample standard RGB videos into high-frequency sequences. These sequences are then processed by the **v2e toolbox** [2], where we explicitly model diverse environments—such as adjusting photon thresholds and noise levels to simulate *low-light* conditions [2]. In parallel, we perform **Subject Localization** by detecting bounding boxes on the original frames. Crucially, since events are asynchronous, we apply **Temporal Interpolation** to the detection masks to align them with the timestamps. Finally, the raw events are spatially cropped and accumulated using these synchronized masks to yield the final event representations.

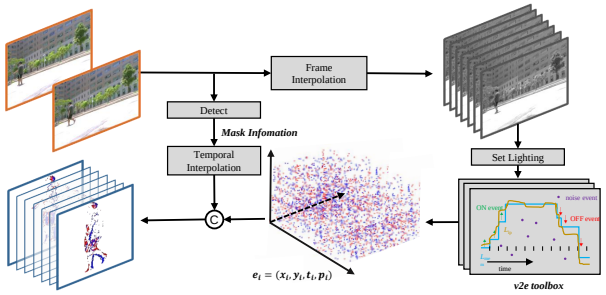


Figure 3. **Overview of the Event Synthesis Pipeline.**

E. Visual Comparison across Modalities

To demonstrate the superiority of event streams under complex illumination, we visualize samples from RGB frames, Silhouettes, and Events in Figure 1. As shown in column (a), all modalities capture clear human bodies under normal light. However, in challenging scenarios such as night (NT) in column (b) or synthesized low-light in column (c), the quality of traditional vision data degrades severely: RGB frames lose most appearance information, and Silhouettes suffer from segmentation failures.

In contrast, **Event streams** rely on brightness changes rather than absolute intensity, which consistently preserve high-fidelity motion outlines and dynamic cues regardless of the lighting conditions, validating their potential for robust all-day gait recognition.

Table 3. Implementation details. The batch size indicates the number of the IDs and the sequences per ID.

DataSet	Batch Size	Schedule	Frames	Steps
SUSTech1K-E	(8, 8)	(20k, 40k, 50k)	30	60k
CCGR-Mini-E	(8, 8)	(20k, 40k, 50k)	30	60k
EV-CASIA-B	(8, 8)	(20k, 40k, 50k)	30	60k

F. Implementation Details

The details of the implementation of the datasets are summarized in Table 3. We use the same training setup across SUSTech1K-E, CCGR-Mini-E, and EV-CASIA-B. Each mini-batch samples 8 identities, with 8 sequences per identity (batch size (8, 8)); each sequence contains 30 frames. Training runs for 60k optimization steps. The learning rate is scheduled by a MultiStepLR scheduler with milestones at 20k, 40k, and 50k steps, where the learning rate is multiplied by gamma = 0.1 at each milestone. Unless otherwise specified, all datasets follow this configuration.

References

[1] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9707–9716, 2023. 2

[2] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1312–1321, 2021. 2