E-S2Feat:Semantic-Guided Spiking Local Feature Detection and Description for Event Cameras

Yang Yi, Juntao Hua, Jinpu Zhang, Liangwei Fan*, Hui Shen** and Dewen Hu

arXiv:2608.14027v1 [cs.CV] 14 Aug 2026

***Abstract*—Benefiting from high temporal resolution and dynamic range, event-based local feature methods have attracted increasing attention. However, event sparsity, noise, and limited texture still hinder robust local feature learning. Deploying such methods on resource-constrained platforms such as unmanned aerial vehicles also requires balancing accuracy and energy efficiency. To address these challenges, this paper proposes E-S2Feat, a spiking neural network framework for event-based local feature detection and description. The framework jointly optimizes local feature learning from the perspectives of feature representation and selection. First, a module-specific spiking activation mechanism preserves fine-grained structural cues and discriminative information under low-bit, energy-efficient inference, thereby improving overall representation fidelity. Furthermore, a semantic-guided feature modulation mechanism leverages semantic priors to refine keypoint response distributions and enhance local descriptor discriminability, thereby guiding the model to extract local features with greater geometric stability and stronger discriminative capability. Experiments on the ECD and EDS datasets show that the proposed method significantly outperforms baseline methods such as SuperEvent in pose estimation accuracy. It also achieves accuracy comparable to its artificial neural network counterpart while delivering an approximately 4.8-fold improvement in theoretical computational energy efficiency. Visual-inertial odometry experiments on the TUM-VIE dataset further verify the effectiveness and practical application potential of the proposed method in complete SLAM systems.**

***Index Terms*—Event cameras, Feature detection, Feature descriptor, SLAM, Semantic segmentation**

I. Introduction

Local feature detection and description are fundamental tasks in computer vision. They provide essential support for downstream applications such as image matching [1], [2], 3D reconstruction [3], [4], and SLAM [5], [6]. Local feature methods based on RGB images [7]–[10] have achieved strong performance in conventional scenarios. However, frame-based cameras are vulnerable to motion blur and exposure failure under high-speed motion or extreme lighting conditions. These effects often degrade feature quality and reduce the robustness of visual localization. In contrast, event cameras [11] offer high temporal resolution, high dynamic range, and low power consumption [12]. They can provide reliable visual information in scenes with rapid motion and challenging illumination. Event cameras have therefore shown strong potential in applications such as autonomous driving and visual SLAM [13], [14]. Nevertheless, event streams are inherently sparse, noisy, and often lack texture, which can degrade the stability and repeatability of extracted features and thus compromise the performance of downstream visual localization.

In practical visual localization systems deployed on resource-constrained platforms such as unmanned aerial vehicles, the front-end feature extractor must not only ensure high matching accuracy but also meet real-time and energy-efficiency requirements, which motivates the exploration of more energy-efficient neural network computation. In recent years, spiking neural networks (SNNs) have attracted increasing attention. SNNs encode information through spike timing and are generally considered more biologically plausible and energy-efficient than conventional artificial neural networks (ANNs). They have been applied to several event-based vision tasks, including classification [15], object detection [16], and optical flow estimation [17]. These studies demonstrate the potential of SNNs for low-power perception and event representation learning. However, existing learning-based methods for event feature detection and description still rely mainly on conventional ANNs architectures. Representative methods such as SuperEvent [18] have demonstrated the effectiveness of architectures such as MaxViT [19] for event-based keypoint detection and description. These methods also achieve strong performance in pose estimation. However, they still rely on continuous-valued activations and conventional deep neural network computation. In comparison, the spiking activation and event-driven computation of SNNs are more compatible with the asynchronous nature of event data. SNNs also support low-bit representations and energy-efficient inference [20]. They therefore offer a promising way to improve the efficiency of event-based visual front-end while preserving feature discriminability.

We propose **E-S2Feat**, the first SNN-based framework for the joint detection and description of local features from event cameras. To improve the representational fidelity of local features, we design module-specific spiking strategies tailored to individual nonlinear modules. These strategies enable low-bit, energy-efficient inference while preserving the fine-grained structural details and discriminative cues essential for keypoint localization and descriptor learning. However, sufficiently faithful feature representation does not necessarily imply that all extracted features can provide stable geometric constraints. Regions such as the sky lack stable and repeatable local structures, where event noise can readily produce spurious patterns and induce unstable feature responses, thereby compromising the reliability of feature matching and pose

Y. Yi, J. Hua, J. Zhang, L. Fan, H. Shen and D. Hu are with the College of Intelligence Science and Technology, National University of Defense Technology, China.

** Principal corresponding author : H. Shen(shenhui@nudt.edu.cn)

* Corresponding author : Liangwei Fan(fanliangwei@nudt.edu.cn)

This work has partially been funded by Research on Spatial Perception and Computing Technologies for Dynamic Complex Scenarios Based on Heterogeneous Brain-inspired Intelligence (U25B2069), Autonomous Learning Theory and Applications of Intelligent Unmanned Systems (T2521006) and Multi-modal Object Tracking in Complex Dynamic Scenarios (62506377).

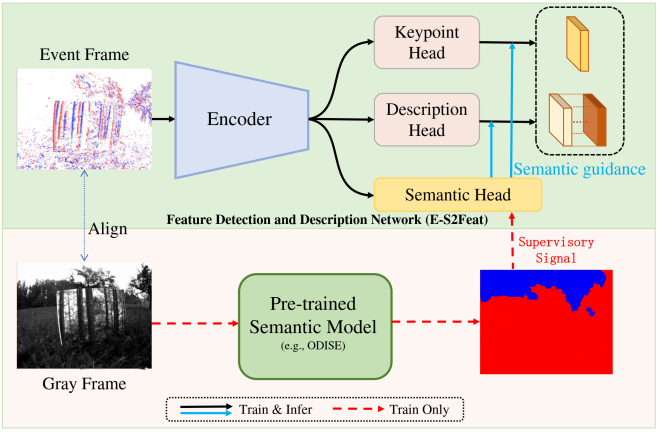

Fig. 1. Overview of the proposed framework. The network is built upon spiking neural networks and incorporates a module-specific spiking activation mechanism, achieving low-bit, energy-efficient inference while preserving feature representation fidelity. Event frames are encoded by a shared spiking backbone and routed to the keypoint, descriptor, and semantic branches. The semantic branch jointly modulates keypoint responses and descriptor features to extract geometrically stable and discriminative local features. It is supervised by a pre-trained semantic model during training, which is unnecessary at inference time.

estimation [12]. To address this issue, we further introduce a semantically guided feature modulation mechanism. To the best of our knowledge, this is the first work to incorporate semantic priors into the joint detection and description of local features from event cameras. The proposed mechanism introduces complementary scene priors at the feature selection level. It exploits semantic information to identify regions that lack stable geometric structures, such as the sky, suppresses unreliable feature responses induced by event noise in these regions, and enhances local descriptor features in geometrically reliable regions. Specifically, we attach a lightweight semantic segmentation head to the shared backbone to predict semantic masks for geometrically unreliable regions. The predicted masks are then used to guide the detection and description branches through score modulation and feature fusion, respectively. Unlike conventional denoising methods that rely on spatio-temporal neighborhood correlations or event statistics [21], [22], the proposed mechanism leverages high-level semantic information to optimize the distribution of keypoint responses and enhance the discriminative capability of local descriptors. In this way, it guides the model toward extracting local features that are more geometrically stable and discriminative.

Consequently, the module-specific spiking strategies and the semantically guided feature modulation mechanism jointly optimize event-based local feature learning at the feature representation and feature selection levels, respectively, together forming a unified framework for event-based local feature detection and description. The overall architecture of E-S2Feat is illustrated in Fig. 1. In summary, the main contributions of this work are as follows:

1) We propose E-S2Feat, the first SNN-based framework for joint local feature detection and description using event cameras. It employs module-specific spiking strategies for different nonlinear components, enabling efficient low-bit inference while largely preserving feature representation capacity.
2) We introduce a semantically guided feature modulation mechanism into event-based local feature detection and description, guiding the model to extract local features that are more geometrically stable and discriminative.
3) Experiments on the ECD, EDS, and TUM-VIE datasets demonstrate that E-S2Feat achieves significant accuracy improvements in pose estimation and Visual-Inertial Odometry (VIO) tasks.

II. Related Works

A. Event-Based Local Feature Extraction

Early research on local feature extraction for event cameras predominantly relied on handcrafted methods, including corner detection based on speed-invariant time surfaces [23], as well as feature tracking and geometric estimation techniques specifically designed for event streams [14], [24], [25]. These approaches are generally characterized by strong interpretability and low computational cost. However, their performance often depends heavily on manually engineered event representations and carefully tuned parameters, which limits their robustness and generalization in the presence of substantial noise, rapid motion, and complex scene variations [12].

In recent years, learning-based approaches to event-based local feature extraction have attracted growing attention. EventPoint [26] and SD2Event [27] jointly learn event-based keypoint detection and description in a self-supervised manner, while LLAK [28] focuses on improving the detection and tracking stability of long-lived keypoints in event streams. More recently, SuperEvent [18] employs cross-modal supervision to learn robust event-based keypoints and descriptors and incorporates them into a keypoint-based SLAM framework, highlighting the potential of learned event local features for visual localization and pose estimation.

To the best of our knowledge, no prior study has explored SNNs for the joint detection and description of event-based local features. This research gap may be attributed to a fundamental challenge: local feature detection and description require the preservation of fine-grained spatial structures and highly discriminative local appearance cues [29], whereas the discrete and capacity-limited spike representations of conventional SNNs can introduce a severe information bottleneck. As a result, it remains difficult to simultaneously achieve the precise spatial localization required for keypoint detection and the rich representational capacity needed for robust feature description.

B. Spiking Neural Networks for Vision

SNNs encode information through the temporal dynamics of discrete spikes. Compared with conventional ANNs, they offer greater biological plausibility and potentially higher energy efficiency, and are widely regarded as a core computational paradigm for neuromorphic computing [30]. The advent of surrogate gradient methods [31] has enabled the end-to-end training of SNNs via backpropagation, substantially accelerating the development of deep SNNs. Nevertheless,

conventional spiking neurons typically communicate through binary spikes [32]–[34], which restricts the amount of information that can be conveyed within a single time step. To compensate for this limited information capacity, high-performance SNNs often rely on multi-timestep rate coding, in which spike responses are accumulated over time to enrich feature representations. This strategy, however, incurs increasing inference latency and computational overhead as the number of simulation timesteps grows. To alleviate this limitation, Fan et al. [35] introduced the Multi-Synaptic Firing (MSF) neuron, which employs multiple firing thresholds to generate multi-level discrete spike responses within a single time step. By increasing the information-carrying capacity of spiking neurons, MSF reduces their reliance on long temporal coding windows.

Despite this progress, multi-level spiking mechanisms such as MSF are primarily tailored to Rectified Linear Unit (ReLU) activations. Smooth nonlinearities widely used in modern visual backbones, including the Gaussian Error Linear Unit (GELU) and the Sigmoid Linear Unit (SiLU), still lack effective spiking counterparts capable of preserving their nonlinear response characteristics while maintaining computational efficiency. Moreover, existing SNNs research has largely focused on high-level visual tasks such as image classification [15] and object detection [16], whereas the potential of SNNs for event-based local feature detection and description remains largely unexplored. To address these limitations, we develop module-specific spiking strategies for heterogeneous nonlinear modules, assigning each module a spiking activation mechanism tailored to its functional characteristics. The proposed design aims to reduce the feature degradation caused by spike quantization while achieving a favorable balance between representation fidelity and computational efficiency.

C. Semantics-Guided Feature Selection

Existing methods commonly leverage semantic segmentation to exclude regions that are unlikely to provide reliable geometric constraints, such as dynamic objects and the sky, thereby reducing mismatches and invalid observations. For example, DynaSLAM [36] and DS-SLAM [37] exploit scene semantics to identify dynamic regions and improve localization robustness in complex environments. Beyond explicit feature filtering, semantic information has also been increasingly incorporated into local feature detection and description. SFD2 [38] introduces semantics-aware constraints to guide the detector toward stable regions, such as buildings and lane markings, while enhancing the discriminative capability of local descriptors. More recently, SAMFeat [39] and GESS [10] have further investigated the use of semantic priors in local feature extraction, description, and matching. Collectively, these studies demonstrate that semantic information can complement low-level textural and geometric cues with high-level scene priors, thereby facilitating the selection of stable features and the learning of more discriminative representations.

However, semantic-guided feature selection remains insufficiently explored for event-based local feature detection and description. Incorporating semantic priors can suppress unreliable feature responses and guide the model to focus on scene regions with stable geometric structures. To this end, this paper integrates a semantic branch into an end-to-end SNNs framework. Built upon the module-specific spiking activation mechanism that improves the representation fidelity of local features, the proposed semantic branch leverages the generated semantic masks to refine keypoint response distributions and enhance the discriminative capability of local descriptors, thereby enabling the extraction of local features with stronger geometric stability and discriminative power.

III. METHODOLOGY

E-S2Feat is an end-to-end SNN-based framework for local feature detection and description with event cameras. The framework jointly optimizes event-based local feature learning at both the feature representation and feature selection levels. At the representation level, the module-specific spiking strategy assigns appropriate spiking activation functions to different network modules according to their functional roles and feature distributions, thereby balancing feature representation capability and computational efficiency. At the selection level, semantic-guided feature modulation exploits scene-level semantic priors to constrain keypoint detection and descriptor generation, improving the geometric reliability and matching stability of local features. The following sections describe the overall architecture, module-specific spiking strategy, semantic-guided feature modulation mechanism, and training strategy.

A. Overall System Architecture

The overall architecture of the proposed system is illustrated in Fig. 2. The input event stream is first converted into a Multi-Channel Time Surfaces (MCTS) representation [18]. Specifically, five approximately logarithmically spaced temporal windows $\Delta t \in \{1, 3, 10, 30, 100\}$ ms and two polarity channels are used to construct a 10-channel multi-scale event representation. This representation preserves both short-term event dynamics and structural information over longer temporal intervals. The MCTS input is then hierarchically encoded by a spiking MaxViT backbone [19]. A feature pyramid network (FPN) further fuses event features across different spatial scales to produce a shared multi-scale representation. The fused features are fed into the detection, description, and semantic prediction heads, which generate a keypoint response heatmap, dense 256-dimensional local descriptors, and a geometric reliability mask, respectively. The semantic predictions guide keypoint detection and descriptor generation through score modulation and feature fusion. Finally, the system outputs semantically constrained keypoint locations and their corresponding local descriptors.

B. Module-Specific Spiking Strategy

A classical approach for converting an artificial neural network into a spiking neural network is to preserve the original topology and weights while replacing continuous activations, such as ReLU, with integrate-and-fire neurons. The

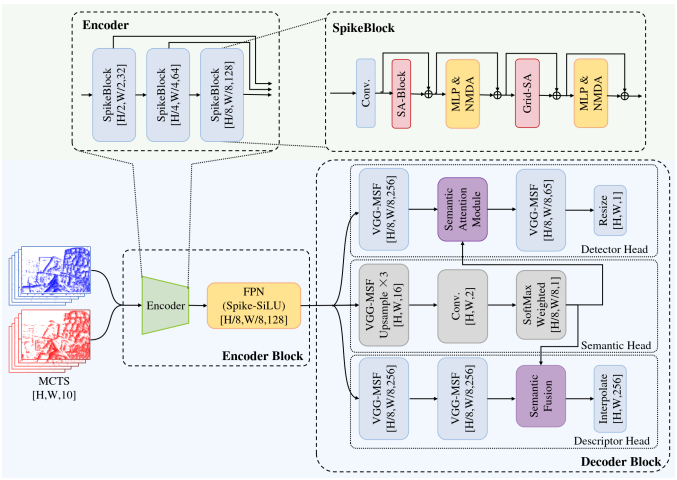

Fig. 2. Overall architecture of E-S2Feat. The spiking MaxViT and FPN extract multi-scale features, while three task branches generate keypoint heatmaps, local descriptors, and a geometric reliability mask. The mask guides feature detection and description through score modulation and feature fusion.

corresponding analog activation values are then approximated by spike firing rates [40]. However, within a limited number of time steps, discrete spike trains cannot precisely represent continuous activations. This mismatch introduces truncation and quantization errors, which may accumulate layer by layer during network propagation [41]. In local feature extraction networks, different modules serve different purposes. Attention modeling, multi-scale feature fusion, and dense semantic prediction have distinct activation distributions, dynamic ranges, and information representation requirements. Previous studies have shown that spike quantization errors are closely related to membrane potential distributions. Applying a uniform spiking activation across different modules may therefore exacerbate activation mismatch and information loss [42].

To this end, we employ module-specific spiking designs for different network components. Specifically, spiking GELU and spiking SiLU are developed to replace the original GELU and SiLU activations, respectively, while ReLU is replaced with the MSF multi-level spiking activation proposed in our previous work [35]. By assigning each module a spiking activation tailored to its functional role and feature distribution, the proposed design preserves the sparsity of spike-based computation while reducing the adverse effects of discrete encoding on local feature detection and description performance.

1) Assignment of Spiking Activations:

a) ***MSF Multi-Level Spiking Activation****:* The detection, description, and semantic heads all require preserving the hierarchy of feature magnitudes within a single time step. However, conventional binary spiking neurons can only represent two states, i.e., spike or no spike, which inevitably leads to excessive compression of feature information. To address this limitation, we replace the ReLU activations in the network with MSF neurons and specifically employ them in the detection, description, and semantic heads to enhance each task branch's capacity to represent fine-grained feature magnitudes. An MSF neuron consists of D synaptic branches with different firing thresholds. For the d-th branch, a binary spike is generated as $s^d = \Theta\left(u_i - V_{\text{th}}^d\right)$ when the membrane potential u_i reaches the corresponding threshold V_{th}^d, where $\Theta(\cdot)$ denotes the Heaviside step function. By aggregating the output spikes from all branches, the MSF neuron can generate an integer spike count ranging from 0 to D within a single time step, thereby providing $D+1$ discrete magnitude levels. The corresponding neuronal dynamics follow those described in [35].

For the l-th ReLU activation, let $x_{\max}^{(l)}$ denote the calibrated upper bound of the corresponding activation values. Given an MSF neuron with D synaptic branches, the quantization step size at this activation is defined as $\Delta_l = x_{\max}^{(l)}/D$. To ensure that the spike count of the MSF neuron corresponds to a round-to-nearest quantization, the firing threshold of the d-th synaptic branch is set to $V_{\text{th}}^{(l,d)} = (d-1/2)\Delta_l$. Consequently, the integer spike count generated by the MSF neuron can be expressed as:

$$n_D(x) = \text{clip}\left(\left\lfloor \frac{x}{\Delta_l} + \frac{1}{2} \right\rfloor, 0, D\right). \tag{1}$$

The corresponding reconstructed output is given by:

$$\widehat{x}_D = \Delta_l \cdot n_D(x) \tag{2}$$

where $\text{clip}(\cdot, 0, D)$ constrains the integer spike count to the range $[0, D]$. For inputs within the calibrated range without upper-bound saturation, i.e., $x \in [0, x_{\max}^{(l)}]$, the reconstruction error of the above round-to-nearest quantization satisfies:

$$|\widehat{x}_D - x| \le \frac{\Delta_l}{2} = \frac{x_{\max}^{(l)}}{2D}. \tag{3}$$

Therefore, with a fixed $x_{\max}^{(l)}$, the upper bound of the quantization error decreases at a rate of $1/D$ as the number of synaptic branches D increases, indicating that a larger number of branches enables a finer approximation of feature magnitudes.

Fan et al. [35] showed that setting $D = 4$ achieves a favorable trade-off between accuracy and energy efficiency for classification and related tasks. Under this configuration, both the integer spike-count range and the representation overhead remain relatively low, making this configuration suitable for tasks with modest requirements on feature magnitude resolution. In contrast, the semantic branch in our framework is required not only to predict regional semantic categories but also to generate pixel-level semantic responses for subsequent keypoint score modulation and descriptor feature fusion. Motivated by these requirements, we uniformly set the number of branches of all MSF neurons to $D = 32$ in our experiments. This configuration substantially reduces the quantization step size, thereby preserving finer-grained spatial responses and feature magnitude hierarchies under single-time-step inference.

b) ***Spiking GELU Activation****:* The attention MLP in MaxViT [19] relies on the GELU activation to perform nonlinear feature transformation along the channel dimension, making its output particularly sensitive to activation approximation errors. To reduce the activation mismatch introduced by spike-based conversion, we propose Spiking GELU, which approximates the GELU response using a continuous basis

expansion with the correct wide-range asymptotic behavior. For an input x, the continuous approximation is defined as

$$Z_{\text{wide}}(x) = w_0 x + c\sqrt{x^2+\delta} + b_0 + \sum_{j=1}^{B} \kappa_j H\left(\frac{x}{\sigma_j}\right) \tag{4}$$

where $H(u) = (1-u^2)\cdot\exp(-u^2/2)$ denotes the Mexican-hat wavelet basis [43], κ_j and σ_j represent the amplitude coefficient and scale parameter of the j-th wavelet basis, respectively, B is the number of wavelet bases, and δ is introduced to smooth the square-root basis function. Accordingly, the asymptotic slopes of the continuous approximation at positive and negative infinity are $w_0 + c$ and $w_0 - c$, respectively. To match the asymptotic behavior of GELU, which approaches the linear function x in the positive region and zero in the negative region, we set $w_0 = c = 0.5$ and constrain b_0 to be close to zero during parameter fitting. Consequently, the linear term together with the even-symmetric square-root term captures the wide-range asymptotic structure of GELU, while the wavelet bases with localized support are employed to fit the residual decay in the intermediate activation range, thereby reducing the tail-slope mismatch caused by finite local basis expansions.

The continuous approximation is subsequently discretized using signed multisynaptic spiking neuron. To this end, we define the symmetric round-to-nearest integer operator as:

$$\mathscr{R}(z) = \text{sgn}(z)\left\lfloor |z| + \frac{1}{2} \right\rfloor. \tag{5}$$

The integer spike output and the corresponding reconstructed output of Spiking GELU are respectively given by:

$$\begin{cases} s_{\text{GELU}}(x) = \text{clip}\left(\mathscr{R}\left(\dfrac{Z_{\text{wide}}(x)}{\gamma}\right), -m_G, M_G\right), \\ \widehat{y}_{\text{GELU}}(x) = \gamma \cdot s_{\text{GELU}}(x) \end{cases} \tag{6}$$

where γ represents the output quantization step size, while M_G and m_G specify the maximum positive and negative integer spike counts of Spiking GELU, respectively. Based on the negative dynamic range of GELU and the calibrated positive output bound $y_{\max}$, they are determined as:

$$\begin{cases} m_G = \max\left(1, \left\lceil \dfrac{\varepsilon_G}{\gamma} \right\rceil\right), \\ M_G = \left\lceil \dfrac{y_{\max}}{\gamma} \right\rceil \end{cases} \tag{7}$$

where $\varepsilon_G \triangleq -\min_x \text{GELU}(x) \approx 0.17$ denotes the magnitude of the minimum value of the ideal GELU function in the negative region. Since the negative responses of GELU have relatively small magnitudes, a limited negative spike range is sufficient to cover its principal negative responses, whereas the positive spike range is determined by the actual activation distribution. In this work, we set the quantization step size to $\gamma = 0.1$, yielding $m_G = 2$ and a negative representation lower bound of -0.2, which fully covers the negative responses of GELU.

*c) **Spiking SiLU Activation**:* The SiLU activation in the FPN convolutional modules is approximated by the proposed Tail-Corrected Wavelet (TCW) formulation for spiking conversion. For an approximation function composed of a linear term, a constant term, and a finite number of wavelet bases, the wavelet components decay to zero as $|x| \to \infty$, causing the tail behavior to be ultimately dominated by the linear term. Therefore, simply increasing the number of wavelet bases, reducing the quantization step size, or expanding the clipping range cannot eliminate the tail errors caused by mismatched asymptotic structures. To preserve the bilateral asymptotic properties of the SiLU function, we adopt the following decomposition:

$$\begin{cases} \text{SiLU}(x) = \text{ReLU}(x) + r(|x|), \\ r(u) = -u \cdot \text{sigmoid}(-u), \quad u \geq 0 \end{cases} \tag{8}$$

where the residual function $r(u) \in [-\varepsilon_S, 0]$ approaches zero as $u \to \infty$, and $\varepsilon_S \triangleq -\min_u r(u) \approx 0.278465$ denotes the magnitude of the minimum negative value of SiLU. In this work, only this bounded decaying residual is approximated using wavelet bases:

$$\widehat{r}(u) = \sum_{j=1}^{B} \kappa_j H\left(\frac{u-\mu_j}{\sigma_j}\right) \tag{9}$$

where μ_j denotes the center of the j-th wavelet basis. The corresponding tail-corrected continuous output is defined as:

$$Z_{\text{TCW}}(x) = \text{ReLU}(x) + \widehat{r}(|x|). \tag{10}$$

No constant bias term is introduced in the residual approximation to ensure $\widehat{r}(u) \to 0$. Meanwhile, the constraint $\widehat{r}(0) = 0$ is imposed during parameter optimization to preserve the value of SiLU at the origin. Accordingly, the wavelet parameters are obtained through the following constrained offline fitting process:

$$\begin{aligned} \min_{\{\kappa_j, \mu_j, \sigma_j\}_{j=1}^{B}} \quad & \int_0^U \left[\widehat{r}(u) + u\,\text{sigmoid}(-u)\right]^2 \mathrm{d}u, \\ \text{s.t.} \quad & \widehat{r}(0) = 0, \qquad \sigma_j > 0. \end{aligned} \tag{11}$$

In this work, a fixed set of $B = 3$ wavelet parameters is fitted offline within the interval $[0, 16]$ and shared across all SiLU activation sites. Therefore, the proposed method does not require storing additional independent fitting parameters for different activation locations. During the discrete spiking stage, the symmetric round-to-nearest integer operator $\mathscr{R}(\cdot)$ defined previously is reused to quantize the continuous output as

$$\begin{cases} s_{\text{SiLU}}(x) = \text{clip}\left(\left\lfloor \dfrac{Z_{\text{TCW}}(x)}{\gamma} + \dfrac{1}{2} \right\rfloor, -m_S, M_S\right), \\ \widehat{y}_{\text{SiLU}}(x) = \gamma \cdot s_{\text{SiLU}}(x) \end{cases} \tag{12}$$

where γ denotes the output quantization step size, and M_S and m_S define the positive and negative integer spike ranges of Spiking SiLU, respectively:

$$\begin{cases} m_S = \max\left(1, \left\lceil \dfrac{\varepsilon_S}{\gamma} \right\rceil\right), \\ M_S = \left\lceil \dfrac{x_{\max}}{\gamma} \right\rceil \end{cases} \tag{13}$$

where $x_{\max}$ represents the calibrated upper bound of the positive output. The negative spike range m_S is determined

according to the theoretically defined minimum magnitude of SiLU, ε_S, rather than adopting the negative range used for GELU. This design covers the negative dynamic range of SiLU and avoids additional saturation errors caused by negative clipping.

2) Single-Step Spike Encoding and Inference: In this work, the network is configured with a single time step. Unlike conventional spike encoding schemes that accumulate information through firing rates over multiple time steps, the input MCTS representation [18] has already explicitly encoded the multi-scale temporal information of the event stream using multiple temporal windows. Therefore, it is unnecessary to further unfold the temporal dimension within the network. Single-step inference avoids the latency and computational overhead associated with multi-step forward propagation, making it better suited to the real-time requirements of local feature extraction from event data. Accordingly, each spiking activation maps continuous features into finite-level discrete responses. After calibration and quantization, the activations are represented as low-bit spike counts or integer states, enabling subsequent linear operations to be implemented as integer- or spike-based accumulations, thereby reducing storage overhead and theoretical computational energy consumption.

In summary, the proposed module-specific spiking activations together with the single-step discrete encoding improve the ability of different functional modules to approximate continuous nonlinearities while effectively controlling inference latency and computational cost. Nevertheless, the limited number of discrete response levels inevitably compresses feature magnitudes and fine-grained discriminative information, whereas increasing either the number of time steps or the response levels incurs additional inference latency and computational overhead. To address this limitation, rather than compensating for information loss solely by improving the precision of spike representations, we further introduce a semantic-guided feature modulation mechanism from the perspective of feature selection. This mechanism exploits semantic information to identify geometrically unreliable regions, such as the sky, and adaptively modulates keypoint responses while enhancing local descriptor features, thereby guiding the network to extract geometrically more stable and discriminative local features. Consequently, the proposed module-specific spiking activations and the semantic-guided mechanism collaboratively optimize event-based local feature learning from the complementary perspectives of feature representation and feature selection, respectively.

C. Semantics-guided Feature Modulation

Visual pose estimation is typically formulated under the assumption of a static and rigid scene, where the inter-frame motion of static feature points is induced primarily by camera motion [36]. However, outdoor scenes often contain several types of geometrically unreliable regions. The sky and nearby distant regions lack stable structures, contain limited texture information, and exhibit high depth uncertainty, making them unsuitable for providing reliable geometric constraints. Dynamic vehicles and pedestrians undergo independent motion and therefore cannot be directly used for camera pose estimation [44]. Consequently, even when events in these regions are triggered by genuine brightness changes, they may still constitute structured interference from the perspective of pose estimation.

Motivated by these observations, we introduce semantic priors into the local feature learning process. For semantic category design, SFD2 [38] employs multi-class semantic priors to assign different weights to different regions. However, because the sparsity of event data makes fine-grained semantic categories difficult to predict reliably, we adopt a binary partition that distinguishes geometrically unreliable regions from reliable background regions and assigns them weights of 0 and 1, respectively. The sky can be identified relatively reliably using cues such as brightness and spatial location. In contrast, dynamic vehicles and pedestrians are more difficult to recognize because grayscale inputs lack color-discriminative cues and the available training data remain limited, leading to a relatively high false-negative rate. Future work will incorporate larger-scale annotations of dynamic scenes to improve dynamic-object recognition and explore finer-grained semantic modeling.

Based on this binary semantic partition, we generate semantic masks for geometrically unreliable regions and use them to guide the detection and descriptor branches through score modulation and feature fusion, respectively. The two mechanisms are described in the following subsections.

1) Keypoint Response Modulation: This part corresponds to the semantic attention module in Fig. 2. Specifically, we first apply spatial rearrangement-based downsampling to the semantic probability map so that its spatial resolution matches that of the detection features. The resulting representation is then processed by two mapping layers, each consisting of a convolution, batch normalization, and ReLU activation, followed by a Sigmoid function to generate a spatial attention map $W_s \in [0,1]^{H/8\times W/8}$. The attention map modulates the detection features in a residual manner:

$$F'_{\text{det}} = F_{\text{det}} \odot \left(1 + \alpha \cdot W_s\right). \tag{14}$$

where F_{det} and F'_{det} denote the detection features before and after modulation, respectively, and $\odot$ represents element-wise multiplication. The learnable scaling parameter α, initialized to 0.5, adaptively controls the strength of semantic modulation.

Under this residual modulation scheme, attention weights in geometrically unreliable regions approach zero, leaving the corresponding feature responses largely unchanged. In contrast, geometrically reliable regions receive larger enhancement weights and therefore produce stronger keypoint responses. During the subsequent Top-K selection, candidate keypoints located in geometrically stable regions are preferentially retained. This encourages the detector to reduce its responses in unreliable regions and focus on scene structures that can provide stable geometric constraints.

2) Descriptor-Semantic Feature Fusion: This part corresponds to the semantic fusion module in Fig. 2. The module adopts a three-branch parallel architecture consisting of a semantic encoding branch, a channel attention branch, and a semantic guidance branch. The semantic encoding branch

maps the input semantic map into a 48-dimensional semantic feature F_{sem} through two 3×3 convolutional layers. The channel attention branch employs the SE mechanism [45]. It first aggregates the global statistics of the descriptor features through global average pooling and then generates the channel attention weights W_{ch} using a two-layer fully connected network with a reduction ratio of 8. The semantic guidance branch maps the semantic feature to a spatial attention weight W_{sp} through a 1×1 convolution, thereby encoding the geometric reliability of different spatial locations. The outputs of the three branches are integrated through joint channel-spatial modulation and residual fusion:

$$\begin{aligned} D_m &= D \odot W_{ch} \odot W_{sp}, \\ D_f &= \text{Conv}_{3\times3}(\text{Concat}[D_m, F_{sem}]), \\ D' &= D + \beta \cdot D_f. \end{aligned} \tag{15}$$

where D and D' denote the original descriptor feature and the semantically enhanced descriptor feature, respectively, and D_m denotes the intermediate feature jointly modulated by the channel and spatial attention weights. The residual scaling coefficient β is set to 0.5 to control the contribution of semantic information during feature fusion. This residual fusion strategy supplements and enhances the original descriptor representation with scene-level semantic information, while preventing the semantic features from directly replacing the local texture representations learned by the backbone.

D. Network Training

1) Supervision Signals: The network is trained end to end under a multi-task supervision framework. The supervision signals are obtained from three sets of pseudo-labels. The pseudo-labels for keypoint detection and descriptor matching are jointly generated by pretrained SuperPoint [8] and SuperGlue [46] using synchronized grayscale frames. Pixel-wise pseudo-labels for the semantic branch are generated by applying the open-vocabulary segmentation model ODISE [47] to the target datasets.

2) Loss Functions: During training, the detection and descriptor branches are supervised by the detection loss L_{det} and descriptor loss L_{desc}, respectively, following the design of SuperEvent [18]. Specifically, L_{det} is an 8×8 grid-based classification cross-entropy loss with an additional dustbin channel, whereas L_{desc} is formulated as a hinge loss. To provide explicit supervision for the newly introduced semantic branch, we further employ a pixel-wise semantic loss L_{sem}. For each pixel (i, j), let p_{ij} denote the predicted probability that the pixel belongs to a geometrically unreliable region, $y_{ij} \in \{0, 1\}$ denote the corresponding pseudo-label, and $|\Omega|$ denote the total number of pixels. The semantic loss is then defined as the pixel-wise binary cross-entropy:

$$L_{\text{sem}} = -\frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \big[y_{ij} \log p_{ij} + (1 - y_{ij}) \log(1 - p_{ij}) \big]. \tag{16}$$

The total loss is a weighted sum of the three terms:

$$L_{\text{total}} = L_{\text{det}} + \lambda_{\text{desc}} L_{\text{desc}} + L_{\text{sem}}. \tag{17}$$

where $\lambda_{\text{desc}} = 10$ is the weight assigned to the descriptor loss.

3) Data Augmentation: During training, event-noise augmentation [12] and homographic augmentation [8] are applied jointly. Event-noise augmentation simulates common sensor effects, including contrast fluctuations, timestamp quantization errors, and defective pixels. Homographic augmentation applies random perspective transformations, scaling, and rotation to both the inputs and labels. This improves the robustness of the model to viewpoint changes.

IV. EXPERIMENTS

To validate the effectiveness of the proposed method, we conduct a comprehensive evaluation from five perspectives: pose estimation, ablation studies, qualitative visualization, VIO Trajectory Evaluation, and inference efficiency.

A. Implementation

The proposed method is implemented in PyTorch. Model training is performed on a single NVIDIA H100 GPU. All evaluation experiments are conducted on a separate workstation equipped with an NVIDIA RTX 4090 GPU. The training data configuration follows SuperEvent [18]. We retain FPV [52], Griffin [53], and MVSEC [54]. DDD20 [55] and ViViD++ [56] in the original configuration are replaced with DSEC [57]. The semantic annotations in DDD20 and ViViD++ are of limited quality and cannot provide reliable supervision for the semantic branch. In contrast, DSEC provides higher-quality semantic annotations. For DSEC, the input is downsampled by a factor of 0.5, and the bottom 10 rows of pixels are masked out to remove sensor-edge artifacts. All inputs are cropped to resolutions divisible by the 8×8 grid size. The network is trained using the Adam optimizer with an initial learning rate of 10^{-3} and a batch size of 12. Training lasts for 10 epochs.

B. Pose Estimation

1) Datasets: The Event Camera Dataset [48] (ECD) is a standard benchmark for event-camera pose estimation. It has a resolution of 180×240 and contains multiple indoor and outdoor sequences with ground-truth 6-DoF camera poses. Event-aided Direct Sparse Odometry [49] (EDS) is a higher-resolution event-camera pose estimation dataset with a resolution of 480×640. It was captured using a DAVIS sensor and contains more richly textured scenes, making pose estimation more challenging.

2) Metrics: Following the evaluation protocol in [18], we use the area under the curve (AUC) of the rotation error at thresholds of 5°, 10°, and 20° as the evaluation metrics. During evaluation, candidate keypoints are selected using a detection threshold of $\tau = 0.01$. Fast Non-Maximum Suppression with a kernel size of 5×5 is then applied to remove duplicate responses. Pose estimation is performed using RANSAC thresholds of $t = 1.0$ for ECD and $t = 3.0$ for EDS.

3) Results: To ensure reliable evaluation, we corrected two potential issues identified in the original evaluation code. We then re-evaluated SuperEvent [18] under the same experimental settings. Tab. I compares the pose estimation accuracy of

TABLE I
POSE ESTIMATION RESULTS ON ECD [48] AND EDS [49] DATASETS.

Methods	ECD: AUC [%]			EDS: AUC [%]		
	@5° ↑	@10° ↑	@20° ↑	@5° ↑	@10° ↑	@20° ↑
RATE [25]	3.3	8.4	18.0	2.1	5.1	10.3
LLAK [28]	0.7	1.4	2.1	0.5	0.7	10.3
SD2Event [50]	16.6	25.3	36.5	11.5	22.1	35.7
STPNet [51]	19.1	31.3	42.9	13.9	24.9	38.2
SuperEvent [18]	22.8	35.8	46.7	25.4	37.5	49.0
Ours (ANNs)	**23.0**	**42.4**	**59.6**	**34.8**	**50.9**	**64.6**
Ours (SNNs)	22.8	42.1	59.4	34.1	50.2	64.0

different methods on the ECD and EDS datasets. Ours (ANNs) denotes the ANNs model equipped with semantic modulation, while Ours (SNNs) denotes its spiking counterpart. On ECD, Ours (ANNs) achieves an AUC@20° of 59.6%, outperforming the re-evaluated baseline by 12.9 percentage points. Ours (SNNs) achieves 59.4%, which is nearly identical to the ANNs version. On the higher-resolution EDS dataset, Ours (ANNs) and Ours (SNNs) achieve AUC@20° scores of 64.6% and 64.0%, respectively. Both models substantially outperform the other methods. Notably, the SNNs version remains comparable to the ANNs version across all metrics. This result shows that the proposed spiking design effectively preserves the model's ability to detect and describe event-based features.

C. Ablation Studies

To evaluate the contribution of each component, we conduct a progressive ablation study starting from the SuperEvent [18] baseline. Components are added sequentially, and the results are reported in Tab. II. Starting from the original SuperEvent [18] training configuration, we first evaluate the effect of introducing the DSEC dataset [57]. The AUC@20° increases from 46.7% to 55.0% on ECD and from 49.0% to 60.8% on EDS. These improvements indicate that training data quality has a substantial effect on event-based feature learning. We then introduce the semantic modulation module. The AUC@20° further increases to 59.6% on ECD and 64.6% on EDS. Finally, we construct the spiking version using the proposed module-specific spiking activations. Compared with the ANNs model, the performance decreases by only 0.2 percentage points on ECD, from 59.6% to 59.4%, and by 0.6 percentage points on EDS, from 64.6% to 64.0%. This minor accuracy loss confirms the effectiveness of the proposed module-specific spiking strategy.

D. Qualitative Visualization

Fig. 3 presents representative outputs of the model's semantic segmentation head, where blue denotes geometrically unreliable regions, such as the sky, and green indicates the retained valid background regions. Even under low-illumination and low-contrast conditions, the SNN-based semantic branch can clearly distinguish between the two types of regions

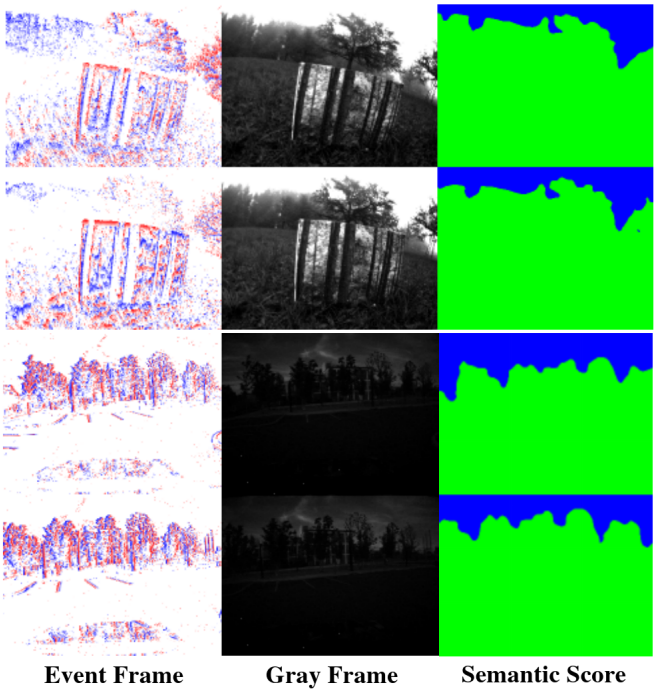

Fig. 3. Visualization of the semantic segmentation output produced by Ours (SNNs). The three columns show the event frame, the grayscale frame, and the prediction from the SNNs semantic branch, respectively. Blue indicates geometrically unreliable regions, such as the sky, while green indicates valid background regions retained for local feature extraction.

and recover their structural boundaries, thereby providing effective semantic guidance for subsequent detection-score modulation and descriptor fusion. Although spike quantization may limit the representation of fine-grained boundary details, the semantic branch equipped with MSF-based multi-level spike activations still achieves prediction quality sufficient for semantic gating. Considering the trade-off among prediction difficulty, computational cost, and semantic guidance effectiveness, we adopt a binary segmentation scheme that distinguishes only between geometrically unreliable regions and valid background regions, rather than performing fine-grained multi-class segmentation. This binary prediction provides adequate geometric priors for feature modulation while maintaining a practical balance between model complexity and

TABLE II
ABLATION EXPERIMENT RESULTS ON ECD [48] AND EDS [49] DATASETS.

Methods	ECD: AUC [%]			EDS: AUC [%]		
	@5° ↑	@10° ↑	@20° ↑	@5° ↑	@10° ↑	@20° ↑
SuperEvent [18]	22.8	35.8	46.7	25.4	37.5	49.0
+ refined training data	21.2	38.9	55.0	33.2	47.8	60.8
+ Semantic	**23.0**	**42.4**	**59.6**	**34.8**	**50.9**	**64.6**
+ Spiking (**Ours SNNs**)	22.8	42.1	59.4	34.1	50.2	64.0

semantic guidance performance.

Fig. 4 further compares the feature detection and matching results of our method with the SuperEvent [18] baseline. The regions highlighted by the yellow boxes reveal clear differences in keypoint distribution between the two methods. The baseline produces numerous responses near the tree–sky boundary, whereas our method employs semantic modulation to enhance detection scores in geometrically stable regions. Consequently, reliable keypoints in these regions are preferentially retained during Top-K selection, while unstable responses near sky boundaries are suppressed. The tree–sky boundary in the upper part of the image further illustrates the difference in matching quality. The baseline produces a considerable number of spatially unstable matches in this region, whereas our method substantially reduces such erroneous correspondences. Meanwhile, within the semantically enhanced regions marked by the green boxes, our method yields denser matches with better spatial consistency, indicating that semantic information guides correspondences toward scene structures with well-defined geometric constraints. These results demonstrate that the performance improvement of our method does not arise from simply increasing the number of keypoints, but rather from the joint enhancement of keypoint localization reliability and descriptor discriminability. At the representation level, spike activations provide low-bit discrete encoding for event features; at the selection level, semantic modulation constrains the spatial distribution of detection responses. Through the coordinated optimization of feature representation and feature selection, the proposed framework enables the network to extract higher-quality keypoints and establish more reliable correspondences.

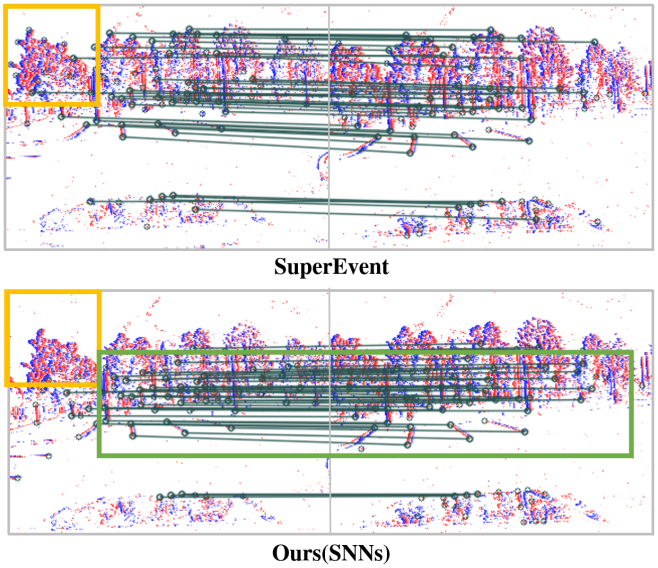

Fig. 4. Qualitative comparison of feature detection and matching between the proposed method and the SuperEvent baseline. Yellow boxes highlight differences in keypoint detection, while green boxes show the improvements in matching density and spatial consistency brought by semantic fusion.

E. VIO Trajectory Evaluation

To further evaluate the practical value of the proposed method in downstream SLAM systems, we integrate the event-based feature methods into the stereo visual-inertial SLAM framework OKVIS2 [58]. We then conduct VIO experiments on the TUM-VIE dataset [59]. Performance is measured by absolute trajectory error (ATE) in centimeters. SuperEvent [18] reports integration results with OKVIS2. However, its interface module has not been released. To ensure a fair comparison, we implement the interface between the event features and OKVIS2. We rerun both SuperEvent and our method using the same pipeline. Therefore, all methods are evaluated under identical conditions.

Tab. III compares the ATE on the small-scale motion-capture sequences of the TUM-VIE dataset [59]. The results of existing event-based methods, including DEIO [60] and ESVO [61], are taken from their original papers. SuperEvent [18] and Ours (SNN) are evaluated using the same VIO pipeline to ensure a consistent comparison. Under this setting, Ours (SNN) achieves the lowest average ATE of 0.96 cm, corresponding to a 44% reduction relative to the reproduced baseline, which obtains an average ATE of 1.71 cm. Fig. 5 presents the trajectory estimation results on the TUM-VIE sequences [59] obtained by integrating OKVIS2 [58] with our method and the baseline features, respectively. The orange solid lines denote the trajectories estimated using our method, the blue solid lines represent those obtained with the baseline, and the gray dashed lines indicate the ground-truth trajectories. As shown, the VIO system based on the proposed spiking features can stably track the camera pose, with the estimated trajectories remaining closely aligned with the ground truth. These results demonstrate that the proposed SNN-based spiking features can be reliably integrated into a complete VIO pipeline.

TABLE III
ATE [CM] ON TUM-VIE MOCAP SEQUENCES [59].

Method	Modality	1d-trans	3d-trans	6dof	desk	desk2	Average
DEIO [60]	Mono E + IMU	1.08	1.12	1.39	1.41	1.19	1.24
ESVO [61]	Stereo E	12.54	17.19	13.46	12.92	4.42	12.11
ES-PTAM [62]	Stereo E	1.05	8.53	10.25	2.50	7.20	5.91
ICRA'24 [63]	Stereo E + IMU	3.85	18.90	failed	8.99	9.47	–
ESVO2 [64]	Stereo E + IMU	3.33	7.26	3.21	6.16	4.02	4.78
OKVIS2 [58] + SuperEvent [18] (our impl.)	Stereo E + IMU	0.49	1.88	2.25	2.07	1.87	1.71
OKVIS2 [58] + Ours (SNNs)	Stereo E + IMU	**0.40**	**1.05**	**1.29**	**1.04**	**1.00**	**0.96**

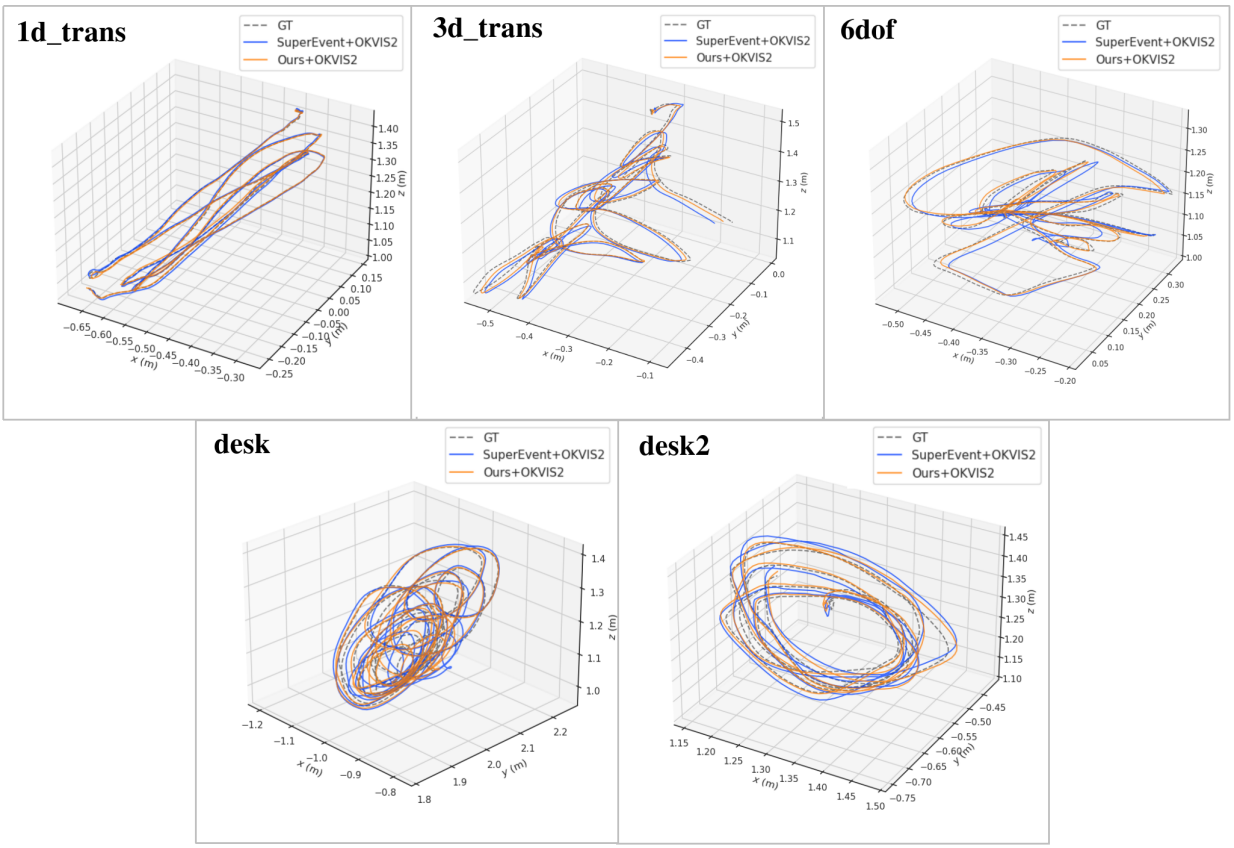

Fig. 5. Trajectory comparison among our method (orange solid lines), the baseline method (blue solid lines), and the ground truth (gray dashed lines) on the TUM-VIE sequences [59].

F. Inference Efficiency

Tab. IV summarizes the model size, inference speed, and theoretical energy consumption of each method. All evaluations are conducted at a resolution of 480×640 using FP32 precision.

Compared with SuperEvent [18], Ours (ANN) introduces a semantic modulation module, increasing the number of parameters from 1.54 M to 2.39 M and reducing the inference speed from 110 FPS to 93 FPS. After further incorporating spiking activations, Ours (SNN) introduces no additional learnable parameters and therefore retains the same parameter count of 2.39 M, while achieving an inference speed of 85 FPS, slightly lower than that of its ANN counterpart. This reduction is mainly attributable to the lack of dedicated hardware support for spiking computation on general-purpose GPUs. Operations involved in spiking activations, such as rounding-based quantization and bit-plane decomposition, must still be simulated using floating-point computation, preventing the model from

TABLE IV
MODEL EFFICIENCY COMPARISON.

Methods	#Params (M)	FPS	Power (mJ)
SuperEvent (ANNs)	1.54	110	7.50
Ours (ANNs)	2.39	93	12.02
Ours (SNNs)	2.39	85	2.52

* All models are evaluated at 480×640 resolution with FP32 precision.

fully exploiting the sparse, event-driven, and asynchronous nature of SNNs. Therefore, the reported inference speed of Ours (SNN) primarily reflects its software simulation efficiency on a conventional GPU and should not be regarded as representative of its actual performance on a target neuromorphic platform.

The theoretical energy consumption is estimated using the operation-count-based method commonly adopted in SNN studies [65]. Owing to the additional computation introduced by the semantic modulation module, Ours (ANN) has a theoretical energy consumption of 12.02 mJ, slightly higher than the 7.50 mJ of the baseline. After spiking activations are introduced, part of the multiply–accumulate operations can be replaced with lower-cost accumulation operations, reducing the theoretical energy consumption of Ours (SNN) to 2.52 mJ. This corresponds to only 33.6% of the baseline energy consumption and 20.9% of that of the corresponding ANN model, yielding an approximately 4.8-fold improvement in theoretical energy efficiency. It should be noted that this estimation is based on standard digital CMOS assumptions and does not account for additional overheads associated with membrane-potential updates and spike communication. Deployment on neuromorphic hardware, such as Loihi, may further improve energy efficiency by exploiting event-triggered asynchronous computation.

In summary, the proposed spiking inference model achieves accuracy comparable to that of its ANN counterpart while providing an approximately 4.8-fold improvement in theoretical energy efficiency, demonstrating the energy-efficiency potential of spiking neural networks for event-based feature extraction and description.

V. Conclusion

This paper proposes E-S2Feat, the first framework to introduce spiking neural networks into event-based local feature detection and description. To accommodate the functional requirements and feature distributions of different network modules, we design spiking GELU and spiking SiLU activations and incorporate the existing MSF multi-level spiking activation. Together, these components form a module-specific spiking strategy that preserves local feature representation while improving inference efficiency. We further propose a semantic-guided feature modulation mechanism. Semantic information is used to adaptively modulate keypoint responses and enhance local descriptors, thereby encouraging the model to extract local features with greater geometric stability and stronger discriminability. Experimental results show that our model achieves competitive performance in pose estimation and visual-inertial odometry. It also maintains accuracy comparable to that of its ANN counterpart while achieving an approximately 4.8-fold improvement in theoretical energy efficiency. These results demonstrate the potential of E-S2Feat as a viable solution for low-power event-based visual localization on resource-constrained platforms. Future work will focus on deploying and optimizing the proposed method on neuromorphic hardware.

References

[1] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk, "Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5173–5182.

[2] C. F. Olson, "Maximum-likelihood image matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 853–857, 2002.

[3] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 8, pp. 1362–1376, 2009.

[4] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6896–6906.

[5] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE transactions on robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[6] Y. Yi, K. Wang, J. Zhang, Z. Tan, X. Wang, H. Shen, and D. Hu, "A plug-and-play learning-based imu bias factor for robust visual-inertial odometry," *IEEE Transactions on Instrumentation and Measurement*, vol. 75, pp. 1–12, 2026.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[8] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[9] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 17 627–17 638.

[10] Y. Yi, X. Chen, J. Zhang, H. Shen, and D. Hu, "Gess: Multi-cue guided local feature learning via geometric and semantic synergy," 2026. [Online]. Available: https://arxiv.org/abs/2604.05359

[11] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128×128 120 db 15μs latency asynchronous temporal contrast vision sensor," *IEEE journal of solid-state circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[12] 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, 2022.

[13] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5419–5427.

[14] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 994–1001, 2018.

[15] M. Yao, H. Gao, G. Zhao, D. Wang, Y. Lin, Z. Yang, and G. Li, "Temporal-wise attention spiking neural networks for event streams classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 221–10 230.

[16] X. Luo, M. Yao, Y. Chou, B. Xu, and G. Li, "Integer-valued training and spike-driven inference spiking neural network for high-performance and energy-efficient object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 253–272.

[17] Z. Wan, Y. Dai, and Y. Mao, "Learning dense and continuous optical flow from an event camera," *IEEE Transactions on Image Processing*, vol. 31, pp. 7237–7251, 2022.

[18] Y. Burkhardt, S. Schaefer, and S. Leutenegger, "Superevent: Cross-modal learning of event-based keypoint detection for slam," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 8918–8928.

[19] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," in *European conference on computer vision*. Springer, 2022, pp. 459–479.

[20] K. Roy, A. Jaiswal, and P. Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, no. 7784, pp. 607–617, 2019.

[21] A. Khodamoradi and R. Kastner, "$o(n)$-space spatiotemporal filter for reducing noise in neuromorphic vision sensors," *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 1, pp. 15–23, 2018.

[22] S. Guo and T. Delbruck, "Low cost and latency event camera background activity denoising," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 785–795, 2022.

[23] J. Manderscheid, A. Sironi, N. Bourdis, D. Migliore, and V. Lepetit, "Speed invariant time surface for learning to detect corner points with event-based cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 245–10 254.

[24] I. Alzugaray and M. Chli, "Asynchronous corner detection and tracking for event cameras in real time," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3177–3184, 2018.

[25] M. Ikura, C. Le Gentil, M. G. Müller, F. Schuler, A. Yamashita, and W. Stürzl, "Rate: Real-time asynchronous feature tracking with event cameras," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 11 662–11 669.

[26] Z. Huang, L. Sun, C. Zhao, S. Li, and S. Su, "Eventpoint: Self-supervised interest point detection and description for event-based camera," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5396–5405.

[27] Y. Gao, Y. Zhu, X. Li, Y. Du, and T. Zhang, "Sd2event:self-supervised learning of dynamic detectors and contextual descriptors for event cameras," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 3055–3064.

[28] P. Chiberre, E. Perot, A. Sironi, and V. Lepetit, "Long-lived accurate keypoints in event streams," *arXiv preprint arXiv:2209.10385*, vol. 1, no. 3, p. 6, 2022.

[29] C. Wang, R. Xu, K. Lu, S. Xu, W. Meng, Y. Zhang, B. Fan, and X. Zhang, "Attention weighted local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 632–10 649, 2023.

[30] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, "Deep learning in spiking neural networks," *Neural networks*, vol. 111, pp. 47–63, 2019.

[31] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate gradient learning in spiking neural networks," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 51–63, 2019.

[32] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of physiology*, vol. 117, no. 4, p. 500, 1952.

[33] A. Shaban, S. S. Bezugam, and M. Suri, "An adaptive threshold neuron for recurrent spiking neural networks with nanodevice hardware implementation," *Nature Communications*, vol. 12, no. 1, p. 4234, 2021.

[34] H. Zheng, Z. Zheng, R. Hu, B. Xiao, Y. Wu, F. Yu, X. Liu, G. Li, and L. Deng, "Temporal dendritic heterogeneity incorporated with spiking neural networks for learning multi-timescale dynamics," *Nature Communications*, vol. 15, no. 1, p. 277, 2024.

[35] L. Fan, H. Shen, X. Lian, Y. Li, M. Yao, G. Li, and D. Hu, "A multisynaptic spiking neuron for simultaneously encoding spatiotemporal dynamics," *Nature Communications*, vol. 16, no. 1, p. 7155, 2025.

[36] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "Dynaslam: Tracking, mapping, and inpainting in dynamic scenes," *IEEE robotics and automation letters*, vol. 3, no. 4, pp. 4076–4083, 2018.

[37] C. Yu, Z. Liu, X.-J. Liu, F. Xie, Y. Yang, Q. Wei, and Q. Fei, "Ds-slam: A semantic visual slam towards dynamic environments," in *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2018, pp. 1168–1174.

[38] F. Xue, I. Budvytis, and R. Cipolla, "Sfd2: Semantic-guided feature detection and description," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5206–5216.

[39] J. Wu, R. Xu, Z. Wood-Doughty, C. Wang, S. Xu, and E. Y. Lam, "Segment anything model is a good teacher for local feature learning," *IEEE Transactions on Image Processing*, 2025.

[40] B. Rueckauer, I.-A. Lungu, Y. Hu, M. Pfeiffer, and S.-C. Liu, "Conversion of continuous-valued deep networks to efficient event-driven networks for image classification," *Frontiers in neuroscience*, vol. 11, p. 682, 2017.

[41] X. Yao, Q. Hu, F. Zhou, T. Liu, G. Li, P. Wang, and J. Cheng, "Towards efficient and accurate spiking neural networks via adaptive bit allocation," *arXiv preprint arXiv:2506.23717*, 2025.

[42] Y. Guo, X. Liu, Y. Chen, L. Zhang, W. Peng, Y. Zhang, X. Huang, and Z. Ma, "Rmp-loss: Regularizing membrane potential distribution for spiking neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 391–17 401.

[43] N. Ricker, "The form and laws of propagation of seismic wavelets," in *World Petroleum Congress*. WPC, 1951, pp. WPC–4039.

[44] M. Altillawi, "Pixselect: Less but reliable pixels for accurate and efficient localization," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 4156–4162.

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[46] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[47] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2955–2966.

[48] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 142–149, 2017.

[49] J. Hidalgo-Carrió, G. Gallego, and D. Scaramuzza, "Event-aided direct sparse odometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5781–5790.

[50] Y. Gao, Y. Zhu, X. Li, Y. Du, and T. Zhang, "Sd2event: Self-supervised learning of dynamic detectors and contextual descriptors for event cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3055–3064.

[51] Y. Zhu, Y. Gao, T. Ding, X. Liu, W. Yang, and T. Zhang, "Spatio-temporal pyramid keypoint detection with event cameras," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[52] J. Delmerico, T. Cieslewski, H. Rebecq, M. Faessler, and D. Scaramuzza, "Are we ready for autonomous drone racing? the uzh-fpv drone racing dataset," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6713–6719.

[53] J. P. Rodríguez-Gómez, R. Tapia, J. L. Paneque, P. Grau, A. G. Eguíluz, J. R. Martínez-de Dios, and A. Ollero, "The griffin perception dataset: Bridging the gap between flapping-wing flight and robotic perception," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1066–1073, 2021.

[54] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3d perception," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2032–2039, 2018.

[55] Y. Hu, J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, "Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–6.

[56] A. J. Lee, Y. Cho, Y.-s. Shin, A. Kim, and H. Myung, "Vivid++: Vision for visibility dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6282–6289, 2022.

[57] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, "Dsec: A stereo event camera dataset for driving scenarios," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4947–4954, 2021.

[58] S. Leutenegger, "Okvis2: Realtime scalable visual-inertial slam with loop closure," *arXiv preprint arXiv:2202.09199*, 2022.

[59] S. Klenk, J. Chui, N. Demmel, and D. Cremers, "Tum-vie: The tum stereo visual-inertial event dataset," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 8601–8608.

[60] W. Guan, F. Lin, P. Chen, and P. Lu, "Deio: Deep event inertial odometry," *arXiv preprint arXiv:2411.03928*, 2024.

[61] Y. Zhou, G. Gallego, and S. Shen, "Event-based stereo visual odometry," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1433–1450, 2021.

[62] S. Ghosh, V. Cavinato, and G. Gallego, "Es-ptam: Event-based stereo parallel tracking and mapping," in *European Conference on Computer Vision (ECCV) Workshops*, 2024.

[63] J. Niu, S. Zhong, and Y. Zhou, "Imu-aided event-based stereo visual odometry," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[64] J. Niu, S. Zhong, X. Lu, S. Shen, G. Gallego, and Y. Zhou, "Esvo2: Direct visual-inertial odometry with stereo event cameras," *arXiv preprint arXiv:2409.02112*, 2024.

[65] S. Kim, S. Park, B. Na, and S. Yoon, "Spiking-yolo: Spiking neural network for energy-efficient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 270–11 277.