Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL35792

Energy-Efficient Event-Based Visual–Inertial Odometry: A Representation–Estimation Coupling Perspective

Jiawen Su

School of Precision Instrument and Opto-electronics Engineering, Tianjin University, Tianjin, China
sjwen136@163.com

Abstract. Event-based visual-inertial odometry (VIO) is often considered a promising sensing solution for low-power micro-robots because event cameras naturally produce sparse, low-latency measurements. However, an efficient sensor does not automatically lead to an energy-efficient system. This paper adopts a system-level view, treating the energy consumption of event-based VIO as an outcome of the interaction between event representation and state-estimation strategy. it examines how information density and computational demand jointly shape system power. From this analysis, it introduces the principle of representation–estimation coupling consistency. Evidence from a broad range of systems supports this principle. The framework also suggests where future low-power VIO systems are likely to converge: event-driven, semi-continuous, and jointly adaptive architectures under strict power budgets. Finally, it provides design guidance for three power ranges and discuss open problems. This study aims to offer a unified perspective for building energy-efficient perception systems for resource-constrained micro-robots.

Keywords: Event-based visual-inertial odometry, Energy efficiency, Representation–estimation co-design, Coupling consistency, Low-power micro-robots

1. Introduction

Micro-robotic platforms, ranging from centimeter-scale vehicles to insect-scale robots, face a fundamental trade-off between the need for low-power perception and the computational cost of accurate state estimation. Under tight size, weight, and power constraints, this conflict quickly becomes a critical system bottleneck. Conventional sensors are not always a good fit: frame cameras generate abundant redundant intensity data, LiDAR is typically too power-hungry, and millimeter-wave radar provides limited angular resolution and sparse measurements. These limitations have encouraged the search for alternative sensing paradigms.

Event cameras report changes in pixel brightness asynchronously, producing a sparse event stream with low redundancy and favorable power characteristics [1]. Yet sensor-level efficiency does not guarantee system-level efficiency; how event data are organized and consumed largely dictates the final energy cost. Inertial measurement units (IMUs) provide high-rate motion constraints at low power, but their estimates drift over time. Combining these complementary strengths gives rise to event-based visual-inertial odometry (VIO) [2].

DOI: 10.54254/2755-2721/2026.GL35792

Most existing work emphasizes accuracy and robustness, while regarding energy efficiency only as a secondary optimization objective [3]. Event representation and temporal estimation are often studied separately. Their combined effect on computation and power has rarely been examined systematically, particularly for micro-robotic platforms with severe resource constraints [4].

This paper examines event-based VIO from the perspective of representation–estimation coupling. Based on a structured review of the literature and a cross-system comparison of reported power characteristics, it develops a system-level framework for understanding energy efficiency in micro-robotic VIO. It also formulates the principle of coupling consistency and uses it to derive practical architectural guidance for embedded systems operating under different power budgets.

2. Event representations: information structure and redundancy

An event representation determines how asynchronous measurements are organized before inference. It shapes the spatiotemporal structure available to later stages, thus directly affecting the information density seen by the estimator [5]. Existing approaches fall into three broad categories according to how they handle time.

Two-dimensional aggregation projects an event stream onto an image plane, discarding most temporal information or retaining it only in a limited form. These representations are simple and allow mature image-processing methods to be reused, but the loss of temporal structure often leads to unnecessary memory traffic and energy consumption. Fixed temporal windows also introduce an inherent trade-off, and the dense frame-like output undermines a key advantage of event cameras: sparse computation. Three-dimensional spatiotemporal volumes reduce this temporal collapse by discretizing both space and time. They preserve ordering across temporal bins and provide richer local motion cues. The main drawback is the rapid growth in representation size: computation and storage scale with voxel count rather than with informative event counts, partly canceling the energy benefit of sparse sensing. Motion-compensated representations avoid a fixed discretization grid. They estimate camera motion and align events into sharper edge structures, reducing data redundancy but adding extra computation.

These three families illustrate a fundamental trade-off between data redundancy and computational burden. None is universally superior; system efficiency depends on the match between the chosen representation and the downstream processing method. For this reason, the estimation paradigm must be considered together with the representation.

3. Estimation paradigms: temporal modeling, complexity, and energy implications

3.1. Discrete-time estimation: structured efficiency with temporal approximation

Discrete-time estimators define system states at fixed timestamps or selected keyframes. Typical examples include extended Kalman filters, particle filters, sliding-window optimization, and factor-graph methods. The multi-state constraint Kalman filter (MSCKF), for example, maintains a window of camera poses and achieves consistency at relatively modest computational cost. Fornasier et al. introduced equivariant filtering into the MSCKF framework to improve convergence and robustness, and DEIO combines differentiable bundle adjustment with IMU preintegration in a factor graph [6, 7]. In these systems, asynchronous events are usually not processed directly. Instead, they are first converted into non-incremental aggregated representations and then treated as measurements aligned with discrete states. This separation simplifies implementation, but it also Imposes a fundamental limitation.

DOI: 10.54254/2755-2721/2026.GL35792

Discrete-time methods benefit from structured updates and predictable runtime. EKF updates can approach linear complexity by exploiting sparsity, whereas sliding-window optimization grows superlinearly with window size and remains more expensive per step even for small windows. The main limitation is the need to approximate asynchronous measurements through temporal discretization. Raising the update rate captures faster event dynamics but increases computation almost proportionally. Aggregating events into dense representations, on the other hand, increases memory traffic and data redundancy. Both choices introduce additional energy cost. Despite inherent latency and sensitivity to update frequency, discrete-time EKF pipelines remain attractive for embedded systems with their simplicity, predictability, and computational tractability.

3.2. Semi-continuous estimation: data-driven updates with lower latency

Semi-continuous estimation retains a discrete state but lets incoming data determine when the estimator updates. Scheduling is data-driven rather than clock-driven. Ribeiro Gomes et al. proposed an event-driven UKF that triggers updates asynchronously upon incoming events while fusing high-rate IMU measurements are fused in parallel. This avoids much of the redundant work caused by fixed-rate updates while maintaining low latency [8].

Semi-continuous methods adapt the update rate to the data rate and thus use resources more efficiently. During periods of low event activity, the estimator performs fewer updates and conserves energy; during rapid motion or bursts of activity, the estimator naturally provides finer temporal resolution. This behavior is useful for micro-robots whose motion is intermittent or highly dynamic. Updating on every single event, however, can create excessive state-propagation overhead. Practical systems therefore tend to use micro-batches, collecting events over a short interval before triggering an update. The event-driven UKF mentioned above uses 1 ms event packets as its smallest processing unit, amortizing state-propagation cost across multiple events [8].

Semi-continuous estimation is well suited to incrementally updated two-dimensional and motion-aware representations. It can preserve temporal sparsity and avoid redundant processing by exploiting the incremental nature of event data. Semi-continuous estimation with progressively refined motion compensation is particularly efficient: each event or micro-batch introduces only a lightweight update to both the representation and the estimator, without requiring full iterative optimization at each step. From an energy perspective, data-driven scheduling reduces idle computation, incremental processing lowers memory overhead, and micro-batching spreads control cost over multiple events. On embedded platforms, this combination often provides a better balance than either conventional discrete-time or fully continuous-time approaches.

3.3. Continuous-time estimation: maximum fidelity at a computational cost

Continuous-time estimators model the trajectory as a continuous function, often using B-splines, Gaussian processes, or continuous-time factor graphs. Measurements can then be incorporated at their exact timestamps, avoiding temporal discretization error and providing a natural formulation for asynchronous sensor fusion. Gao et al. proposed CB-VIO, which tracks event features in a continuous spatiotemporal window using cubic B-splines and fuses them with IMU data [9]. Chui et al. optimized a B-spline trajectory for event-based feature tracking in a sliding-window framework [10]. Wang et al. introduced AsynEIO, which fuses asynchronous event and inertial measurements through Gaussian-process regression [11].

The main drawback of continuous-time estimation is its computational cost. The core problem is usually a nonlinear optimization over trajectory parameters. A full Gaussian-process formulation can

DOI: 10.54254/2755-2721/2026.GL35792

grow cubically with the number of measurements unless sparsification is used. B-spline methods benefit from local support, but still require repeated evaluation of basis functions and Jacobians. Factor-graph formulations involve large sparse optimization problems. Even these costs often exceed the capabilities of low-power micro-robots. Continuous-time estimators pair naturally with motion-compensated or three-dimensional spatiotemporal representations, but dense representations combined with batch optimization can impose high computation and memory demands, making real-time operation difficult. Nonlinear solvers also introduce noticeable latency, sometimes reaching hundreds of milliseconds in batch settings. These methods are computationally intensive and dominated by floating-point operations. Without dedicated hardware or substantial optimization, they are rarely compatible with strict micro-robot power budgets [12].

4. System-level energy analysis through representation–estimation co-design

4.1. Representation and estimation as the main sources of system energy use

From an information-processing perspective, a complete event-based VIO pipeline comprises sensing, representation, and estimation layers. The sensing layer converts physical signals into digital measurements. The representation layer structures the asynchronous event stream, and the estimation layer uses the encoded information to propagate and optimize the system state. All three layers consume energy, but the sensing cost is largely fixed by the hardware. In practice, most of the design freedom and much of the system energy consumption lie in computation and memory access within the representation and estimation layers.

Representation and estimation are distinct yet complementary. The representation determines the information density presented to the estimator, while the estimator determines how effectively that information is used. Their interplay defines the system trade-off among accuracy, latency, and power. Their relationship is bidirectional: the information density selected by the front end limits the amount of useful computation available to the back end, while the estimator imposes requirements on the structure and timing of the representation. Total system energy is not simply the sum of two independent modules. It depends strongly on the degree of match between the two.

4.2. Power patterns in existing event-based VIO systems

In ultra-low-power settings below 1 W, LEVIO achieves real-time localization on a RISC-V SoC, consuming under 100 mW and outputting poses at 20 Hz [13]. Its CCMAX front end reduces floating-point operations by 42% and energy consumption by 87% [14]. Both designs deliberately avoid high-dimensional dense voxel representations and pair lightweight representations with filtering-based estimators. Their information density and computational scale are reduced together.

When power is no longer a hard constraint — systems often sacrifice efficiency for accuracy. The system of Tejero Ruiz runs in real time on a Jetson Nano and consumes about 15 W [15]. Within the intermediate range of roughly 1–5 W, most systems strike a balance between accuracy and computational cost. MA-EVIO reports a mean pose error of 0.19 on DAVIS240c [16]. TrinitySLAM tightly couples events, grayscale frames, and IMU measurements, yielding 28% higher accuracy, half the latency, and 2.2×lower energy consumption [17]. ESS-MSCKF uses a stereo event-camera front end with a filtering back end to avoid repeated batch optimization [18]. Elamin keep batch latency within 5–10 ms while reporting an accuracy improvement of about 50% [19]. Designs in this power limit representation density and estimator complexity.

DOI: 10.54254/2755-2721/2026.GL35792

Both SpikeVPR and the work of Tenzin use spiking neural networks on neuromorphic hardware and therefore belong to a different design space; they are not included in direct comparisons with conventional platforms [20, 21]. eKalibr-Inertial performs high-accuracy spatiotemporal calibration through batch continuous-time optimization [22]. Systems of this kind rely on dense voxels or heavily accumulated event representations together with continuous-time batch optimization.

4.3. Representation–estimation coupling consistency: a design principle for low-power event-based VIO

Energy optimization in event-based VIO is fundamentally a co-design problem. High system power rarely stems solely from a rich representation or a complex estimator; rather, it often arises from a mismatch between the amount of information produced by the front end and the amount of computation the back end can use effectively. One type of mismatch occurs when the representation is heavily compressed but the estimator remains expensive. If events have already been reduced to sparse event frames, adding a batch optimizer offers little accuracy gain, because the missing temporal detail cannot be recovered through more iterations. The extra computation simply consumes energy. The opposite mismatch occurs when a dense voxel grid or detailed motion-compensated representation is paired with a simple filter. The front end incurs substantial storage and memory-transfer cost, but the lightweight estimator cannot fully exploit the retained motion information. In both cases, cost is spent without a proportional gain in accuracy. Efficient design prioritizes coordinated representation and estimator scaling under fixed power budgets, rather than separately minimizing representation volume or estimator complexity. This balanced correspondence is defined as representation–estimation coupling consistency. Well-designed low-power VIO systems maintain comparable representation density and estimator complexity; deviating from this equilibrium results in resource waste or degraded performance.

This analysis yields three practical principles. First, the system power budget directly limits the feasible representation space. Encoding should therefore begin with the energy constraint rather than from an unconstrained goal of preserving every available feature. Second, back-end complexity should match the information density produced by the front end. When a representation is too weak to support sophisticated optimization, additional computation brings little benefit; when it is richer than the estimator can use, information is wasted and memory traffic increases. Third, under tight power constraints, preserving the essential temporal structure with a moderately simplified estimator is often more effective than aggressively compressing the representation while retaining an expensive back end. Once motion detail is removed at the representation stage, later algorithms cannot reconstruct it reliably.

The main energy bottleneck in event-based VIO therefore hinges on the match between representation and estimation complexity. Efficient systems preserve only the temporal structure that matters and process it at a rate consistent with its information density. For micro-robots, the best design lies in the joint space of representation and estimation rather than in optimizing either one alone. Because event cameras naturally produce sparse measurements, they offer considerable room for low-power co-design. Future systems should adapt representation and estimation jointly under the combined constraints of task demand, power budget, and hardware capability.

5. Conclusion

This paper treats energy efficiency in event-based VIO as a system property emerging from the interplay between event representation and estimation strategy. Organizing existing methods along

DOI: 10.54254/2755-2721/2026.GL35792

these two dimensions provides a basis for comparison. The analysis suggests that efficiency is determined less by the absolute sparsity of the event stream than by the preservation and exploitation of that sparsity throughout the processing pipeline. This work proposes representation–estimation coupling consistency to describe the alignment between information density and computational response. Efficient systems show the same pattern: sparse or motion-compensated representations are paired with incremental or event-driven estimators, while dense representations are reserved for high-performance settings with looser power constraints. Reported systems ranging from sub-100 mW embedded platforms to multi-watt optimized pipelines support this observation.

The framework also offers guidance for future system design. Under strict power limits, viable architectures are likely to move toward event-driven pipelines and semi-continuous estimation. Two limitations remain. First, coupling consistency remains qualitative; developing a quantitative metric is an important direction for future work. Second, hardware hierarchy, data movement, and parallel execution can dominate real energy consumption, but these effects are not modeled explicitly here.

Overall, this paper frames energy efficiency in event-based VIO as a joint information-and-computation design problem. An effective system should not optimize individual modules in isolation, but instead coordinate representation and estimation. The key is to ensure that information is processed at the same scale at which it is produced. Mismatch, rather than complexity alone, is the deeper source of inefficiency.

References

[1] Gallego, G., Delbrück, T., Orchard, G., et al. (2022). Event based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(1), 154–180.

[2] Huang, K., Zhang, S., Zhang, J., & Tao, D. (2024). Event based simultaneous localization and mapping: A comprehensive survey. *arXiv*. https://arxiv.org/abs/2304.09793

[3] Mahlknecht, F., Gehrig, D., Nash, J., et al. (2022). Exploring event camera based odometry for planetary robots. *IEEE Robotics and Automation Letters, 7*(4), 8651–8658.

[4] Zhang, J., Yu, X., Sier, H., et al. (2024). Event based sensor fusion and application on odometry: A survey. *arXiv*.

[5] Zheng, X., Liu, Y., Lu, T., Hua, T., Pan, W., Zhang, D., & Wang, L. (2023). Deep learning for event based vision: A comprehensive survey and benchmarks. *arXiv*. https://arxiv.org/abs/2302.08890

[6] Fornasier, A., van Goor, P., Allak, E., Mahony, R., & Weiss, S. (2024). MSCEqF: A multi state constraint equivariant filter for vision aided inertial navigation. *IEEE Robotics and Automation Letters, 9*(1), 731–738.

[7] Guan, W., Lin, F., Chen, P., & Lu, P. (2025). DEIO: Deep event inertial odometry. In *IEEE/CVF International Conference on Computer Vision Workshops*.

[8] Ribeiro Gomes, J., Gaspar, J., & Bernardino, A. (2023). Event based feature tracking in a visual inertial odometry framework. *Frontiers in Robotics and AI, 10*.

[9] Gao, X., Li, Z., Liu, H., Xue, H., Chen, B., & Ge, S. S. (2023). Cubic B-spline based feature tracking for visual–inertial odometry. *IEEE Transactions on Instrumentation and Measurement, 72*.

[10] Chui, J., Klenk, S., & Cremers, D. (2021). Event based feature tracking in continuous time with sliding window optimization. *arXiv*. https://arxiv.org/abs/2107.04536

[11] Wang, Z., Liu, Y., Xu, J., Huang, P., & Li, X. (2025). AsynEIO: Asynchronous monocular event inertial odometry using Gaussian process regression. *IEEE Transactions on Robotics, 41*, 5020–5039.

[12] Scherer, M., Mayer, P., Di Mauro, A., Magno, M., & Benini, L. (2021). Towards always on event based cameras for long lasting battery operated smart sensor nodes. In *IEEE International Instrumentation and Measurement Technology Conference*.

[13] Kühne, J., Vogt, C., Magno, M., & Benini, L. (2026). LEVIO: Lightweight embedded visual inertial odometry for resource constrained devices. *IEEE Sensors Journal, 26*(3).

[14] Min, K., Choi, J., & Lee, W. (2026). Coarse to fine contrast maximization for energy efficient motion estimation in edge deployed event based SLAM. *Micromachines, 17*(2), 176.

[15] Tejero Ruiz, D., Solís Martín, D., Pérez Grau, F. J., & Borrego Díaz, J. (2026). Indoor UAV navigation using event cameras and intermediate frame reconstruction. *Computer Vision and Image Understanding*.

DOI: 10.54254/2755-2721/2026.GL35792

[16] Shahraki, M., Elamin, A., & El Rabbany, A. (2025). MA EVIO: A motion aware approach to event based visual–inertial odometry. *Sensors, 25*(23), 7381.
[17] Cai, X., Xu, J., Deng, K., Lan, H., Wu, Y., Zhuge, X., & Yang, Z. (2024). TrinitySLAM: On board real time event image fusion SLAM system for drones. *ACM Transactions on Sensor Networks, 20*(6).
[18] Keskin, A. (2025). *Event camera based pose estimation in stereo visual inertial odometry* [Master's thesis]. Middle East Technical University.
[19] Elamin, A., Jacob, S., & El Rabbany, A. (2025). Event based visual/inertial odometry for UAV indoor navigation. *Sensors, 25*(1), 61.
[20] Keime, G., Cuperlier, N., & Cottereau, B. R. (2026). Event driven neuromorphic vision enables energy efficient visual place recognition. *HAL*. https://hal.science/hal-05573824
[21] Tenzin, S. (2026). *Towards neuromorphic visual SLAM: A spiking neural network for efficient pose estimation and loop closure* [Master's thesis]. Edith Cowan University.
[22] Chen, S., Li, X., & Yuan, L. (2025). eKalibr Inertial: Continuous time spatiotemporal calibration for event based visual inertial systems. *arXiv*. https://arxiv.org/abs/2509.05923