Algorithms and Intelligence

Vol 2 Issue 1 (2026)

www.algointel.org

Event-stream Dynamics Priors for Low-latency Visual Representation Learning

Nicholas Ho[1], Ka-Lun Shum[2]

[1] Department of Computer Science, Faculty of Science and Engineering, Hong Kong Chu Hai College, Hong Kong, Hong Kong SAR, China

[2] Department of Computer Science and Technology, Tsinghua University, School of Aerospace Engineering, Tsinghua University, Beijing, China

Abstract

The rapid evolution of neuromorphic engineering has introduced event-based cameras as a revolutionary alternative to traditional frame-based visual sensors. Unlike standard cameras that capture full images at fixed intervals, event cameras respond asynchronously to logarithmic changes in per-pixel illumination, yielding a sparse, continuous-time data stream with microsecond temporal resolution and exceptionally high dynamic range. Despite these hardware advantages, extracting meaningful representations from such asynchronous streams with minimal latency remains a profound challenge for modern computer vision architectures. Standard deep learning frameworks, optimized for dense, synchronous tensor operations, inherently compromise the latency benefits of event streams by requiring artificial temporal accumulation or framing. This paper introduces the concept of event-stream dynamics priors as a fundamental inductive bias to guide low-latency visual representation learning. By embedding physical kinematic principles directly into the architecture and optimization objectives of continuous-time neural networks, we enable the model to infer latent spatiotemporal states instantly upon the arrival of individual events. We present a comprehensive theoretical framework for these dynamics priors, detailing their integration into asynchronous learning pipelines. Extensive experimental analysis across multiple neuromorphic datasets demonstrates that integrating dynamics priors significantly reduces the time-to-accuracy metric, allowing networks to converge to robust representations at sub-millisecond latencies. The findings suggest a paradigm shift in processing asynchronous visual data, moving away from purely data-driven feature extraction towards physics-informed continuous representation learning.

Keywords

Neuromorphic vision, Representation learning, Event-stream processing, Continuous-time networks.

1. INTRODUCTION

1.1 The Neuromorphic Paradigm Shift

The transition from synchronous, frame-based visual acquisition to asynchronous, event-driven sensing represents one of the most significant paradigm shifts in modern computer vision. Traditional image sensors, such as charge-coupled devices and active pixel sensors, operate on a rigid global shutter or rolling shutter mechanism, acquiring information across the entire pixel array at predefined intervals. This synchronous methodology, while highly successful in producing easily interpretable dense matrices for static analysis, suffers from inherent limitations in highly dynamic environments. Standard cameras are bound by the frame rate, leading to motion blur during rapid scene changes and significant latency bottlenecks between the occurrence of a physical event and its digitization. Furthermore, they generate massive amounts of redundant data when observing static scenes, resulting in excessive power consumption and computational overhead. In stark contrast, event-based cameras, inspired by the biological functioning of the mammalian retina, operate on a completely different principle. These sensors contain independent pixels that monitor the log-intensity of incoming light continuously. A pixel only triggers an output, termed an event, when the local change in brightness exceeds a predefined threshold. This localized, asynchronous operation means that the sensor output is driven purely by the dynamics of the scene. The literature also considers pixel-wise calibration for a multi-focus microscopic 3D imaging system [1]. Another relevant line of investigation concerns empowering large language models to progressively understand and generate 3D mesh [2].

1.2 The Challenge of Low-latency Processing

The primary appeal of event cameras lies in their microsecond temporal resolution, high dynamic range, and minimal redundant data generation. However, realizing the theoretical latency advantages of these sensors in practical downstream tasks, such as object detection, tracking, or autonomous navigation, introduces profound computational challenges. The output of an event camera is a non-uniform, sparse stream of coordinates and timestamps, lacking the dense spatial structure upon which decades of computer vision algorithms have been built. Early approaches to event-stream processing attempted to bridge this gap by accumulating events over fixed time windows or fixed event counts to construct pseudo-frames or voxel grids. While this accumulation allows for the direct application of highly optimized convolutional neural networks, it inherently destroys the continuous-time nature of the data. Accumulation introduces artificial latency proportional to the window size, completely negating the microsecond responsiveness of the sensor. Alternatively, purely asynchronous architectures, such as spiking neural networks, offer a more biologically plausible processing route but have historically struggled with complex representation learning due to the non-differentiable nature of discrete spike communication. Bridging the gap between the raw, asynchronous event stream and high-level semantic representations without introducing accumulation latency remains an open problem in the field. Li and colleagues examine multi-stream graph convolution network for driver head pose estimation [3]. Related work has addressed vision-Language Model-Guided Uncertainty-Aware Cross-Modal sensor fusion for robust autonomous driving [4].

1.3 Introducing Dynamics Priors

To overcome the latency bottleneck induced by spatial accumulation, this paper proposes the integration of event-stream dynamics priors into the representation learning process. The core hypothesis is that the generation of events is fundamentally governed by the physical kinematics of the scene and the sensor. Because events are exclusively generated by relative motion between the camera and the environment, the resulting spatiotemporal point cloud contains implicit, highly structured information regarding scene depth, velocity, and object boundaries. By formulating

these physical constraints as inductive biases, or dynamics priors, we can guide the learning process of continuous-time neural networks. A dynamics prior acts as a regularization mechanism that expects the latent representations to evolve smoothly and predictably according to physical laws, rather than relying solely on the data-driven optimization of isolated features. This allows the network to anticipate future states based on historical trajectories, enabling the instant readout of robust representations upon the arrival of single events without waiting for a dense spatial pattern to form. The integration of such priors shifts the burden of temporal integration from the input data processing stage to the internal latent dynamics of the network itself. Cong and colleagues examine weakly supervised 3d multi-person pose estimation for large-scale scenes based onmonocular camera and single lidar [5].

2. Background and Related Work

2.1 Evolution of Event-based Vision Sensors

The conceptualization of event-driven vision sensors dates back to the development of the silicon retina, which sought to emulate the analog processing capabilities of biological photoreceptors. Modern Dynamic Vision Sensors have refined this concept, achieving widespread adoption in robotics and edge computing domains due to their microsecond precision and robustness to extreme lighting conditions. Unlike traditional sensors that suffer from under-exposure or over-exposure in high dynamic range scenes, the independent logarithmic response of event pixels allows for simultaneous observation of regions with vastly different illuminations. The asynchronous nature of the output means that the temporal resolution is virtually continuous, limited only by the bandwidth of the digital readout circuitry. However, this unique data format necessitates novel algorithmic approaches. Researchers have extensively explored how to interface these sensors with conventional digital processors, leading to the development of specialized neuromorphic processors and asynchronous logic circuits tailored for event streams. Despite hardware advancements, the software ecosystem for extracting high-level semantic meaning from these streams is still heavily dependent on synchronous approximations. Meng and colleagues examine counterfactual audit and repair for Trust-Aware Multi-Signal streaming LiDAR segmentation [6]. Chang Liu examines emotion-aware diffusion for editable 3D Gaussian talking head [7].

2.2 Standard Approaches to Spatiotemporal Representation Learning

The dominant methodology for learning visual representations from event streams currently involves mapping the asynchronous data into synchronous tensor formats. Time-surface formulations, for example, encode the timestamp of the most recent event at each pixel, creating a dense map of historical activity that can be processed by standard spatial convolutions. While computationally efficient, time surfaces discard the fine-grained temporal ordering of multiple events occurring at the same pixel within a short window. More advanced techniques utilize spatiotemporal voxel grids, which discretize the event stream in both space and time, preserving more information at the cost of significantly higher memory consumption and processing latency. Graph neural networks have also been proposed to treat the event stream as a dynamic point cloud, constructing spatial neighborhoods based on Euclidean distance and temporal proximity. While graph-based approaches preserve the sparsity of the data, the continuous updating of graph topology as new events arrive is computationally expensive and scales poorly to the millions of events per second generated in complex environments. All these standard approaches fundamentally treat time as an additional spatial dimension, discretizing it in a way that introduces latency and computational bottlenecks. Another relevant line of investigation concerns conditional denoising diffusion probabilistic model for point cloud upsampling [8].

2.3 The Role of Inductive Biases and Priors

In deep learning, an inductive bias refers to the set of assumptions a learning algorithm uses to predict outputs for inputs it has not encountered. Convolutional neural networks, for instance, possess a strong inductive bias towards translation invariance and spatial locality. When dealing with temporal data, recurrent neural networks assume sequential dependence. In the context of event-based vision, relying solely on generic spatial or sequential biases is insufficient due to the unique genesis mechanism of the data. Events are not arbitrary points in space and time; they are the direct consequence of brightness constancy equations under optical flow. Introducing priors based on this physical reality allows networks to learn more efficiently from less data. Previous work in physical machine learning has demonstrated that embedding differential equations governing fluid dynamics or structural mechanics into neural network architectures significantly improves generalization. By adapting this philosophy to event streams, we can formulate dynamics priors that constrain the learned latent space to adhere to kinematic continuity. This prevents the network from generating erratic representations when faced with sparse inputs and facilitates the stable extrapolation of features between consecutive events. Another relevant line of investigation concerns accurate and convenient method of vehicle spatiotemporal distribution recognition based on computer vision [9]. Related work has addressed structural pruning of large vision language models: A comprehensive study on pruning dynamics, recovery, and data efficiency [10].

3. Theoretical Foundations of Event-Stream Dynamics

3.1 Mathematical Formulation of Asynchronous Event Streams

To rigorously define the concept of dynamics priors, we must first establish a mathematical framework for the event stream itself. Two-stream action recognition separates spatial appearance from temporal motion evidence, a useful reference for evaluating what each visual stream contributes [11]. An event camera outputs a sequence of discrete occurrences triggered by relative illumination changes. Let the spatial domain of the sensor be defined as a two-dimensional grid of pixels. An event is generated at a specific pixel location at a precise continuous time when the logarithmic brightness changes by a set threshold magnitude. This can be formally represented as a tuple comprising the spatial coordinates, the timestamp, and the polarity of the change, indicating whether the brightness increased or decreased. The entire event stream over a given observation period is therefore a set of these asynchronous tuples. Because the temporal dimension is strictly monotonically increasing and not constrained to a fixed grid, traditional sampling theory and Nyquist-Shannon reconstruction paradigms do not directly apply. The generation of these events is described by the linearized event generation model, which relates the temporal derivative of the log-intensity image to the spatial gradient of the image and the optical flow vector at that point. This fundamental relationship is the cornerstone of our proposed dynamics priors, as it links the discrete event occurrences directly to the continuous motion of the scene. The literature also considers learning Event-to-Panorama image reconstruction [12]. Related work has addressed object fidelity diffusion for remote sensing image generation [13].

3.2 Formulating the Kinematic Dynamics Prior

The objective of low-latency representation learning is to maintain a continuous, high-dimensional latent state that can be queried at any arbitrary time to yield meaningful semantic information. To achieve this without spatial accumulation, the latent state must evolve continuously between the arrivals of individual events. We propose modeling this evolution using a kinematic dynamics prior. This prior assumes that the latent features correspond to physical attributes of the scene that move and deform according to smooth vector fields. We define the evolution of the latent representation matrix as a partial differential equation governed by a learned velocity field. Another relevant line of investigation concerns zero-Shot unified image restoration via latent diffusion recurrent posterior sampling [14]. Another relevant line of investigation concerns

cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose [15]. To enforce this structural constraint during the training phase, we construct a spatiotemporal optimization objective that penalizes deviations from kinematic continuity. The total loss function incorporates both a task-specific empirical risk term and the dynamics prior regularization term. The regularization term is designed to minimize the residual of the continuity equation in the latent space. We express this mathematically as follows: Another relevant line of investigation concerns inferring 3d hands from both text prompts and body dynamics [16]. Another relevant line of investigation concerns singing timbre popularity assessment based on multimodal large foundation model [17].

$$L_{total} = \sum_{i=1}^{N} L_{task}(f(x_i, t_i), y_i) + \lambda \int_{\Omega} \int_{0}^{T} (\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v))^2 dx dt$$

In this formulation, the first term represents the standard loss evaluated at the discrete event times, where the function extracts the representation and compares it to the target label. Contrasting cluster assignments provides a self-supervised route to visual representation learning, making cluster stability an architectural as well as an evaluation question [18]. The second term, weighted by the hyperparameter lambda, is the dynamics prior. It enforces that the density of the latent features and their associated velocity field adhere to a continuous transport equation over the spatial domain and the entire time horizon. The velocity field is implicitly learned by the network alongside the features. By minimizing this integral, the network is heavily penalized for disjointed jumps in the latent representation, forcing it to learn a smooth, predictive internal model of the scene dynamics. For a broader comparison, related research addresses scene-Aware enhanced Edge-Cloud collaborative industrial vision inspection with multimodal LLM [19].

3.3 Implications for Latency and State Update

The inclusion of the continuous transport equation in the loss function fundamentally alters the learning dynamics. Without the prior, a neural network processing sequential events must wait until a statistically significant number of events have accumulated in a local spatial region to confidently update its feature maps. The prior, however, acts as a predictive engine. When an event updates the latent state at a specific coordinate, the learned velocity field dictates how that update propagates through space over time, even in the absence of new incoming events in adjacent pixels. This means that the entire latent representation remains active and synchronized with the physical world. When a downstream task requires an output, the representation is instantaneously available and robust, requiring zero accumulation delay. The latency is therefore reduced from the macro-scale of frame accumulation to the micro-scale of a single forward pass of the continuous-time network architecture. Related work has addressed traversal-Free ParallelFiltering and adaptive shrinkingfor Large-Scale LoD 3D GaussianSplatting [20]. Another relevant line of investigation concerns intelligent End–Edge–Cloud architecture for visual IoT-Assisted healthcare systems [21].

4. Methodology for Representation Learning

4.1 Prior-infused Network Architecture

Translating the theoretical continuous-time dynamics prior into a computable neural network architecture requires careful design choices. We employ Continuous-Time Recurrent Neural Networks as the foundational building blocks. Unlike standard recurrent units that update their hidden state in discrete steps based on a clock, Continuous-Time Recurrent Neural Networks define the hidden state derivative as a parameterized function of the current state and the incoming input. To inject the dynamics prior, we construct a specialized hidden layer where the state evolution is explicitly split into two distinct pathways: a data-driven update pathway triggered asynchronously by incoming events, and a prior-driven autonomous evolution pathway governed by the learned velocity field. The autonomous pathway utilizes neural ordinary differential equation solvers to continuously integrate the state forward in time between event arrivals. This

ensures that the representation is never stale. When a new event arrives, its spatial coordinates are used to locally perturb the continuous state through a spatial attention mechanism, ensuring that the network remains grounded in the actual sensor observations while relying on the prior for temporal extrapolation. The literature also considers empower vision applications with LoRA LMM [22]. Chen and colleagues examine hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception [23].

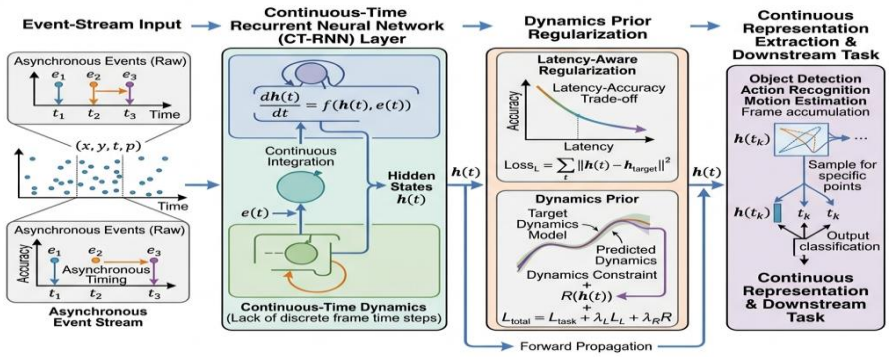


Figure 1Comprehensive Schematic of Latency

4.2 Latency-aware Optimization Strategy

Training a network to optimize both accuracy and latency requires a specialized learning schedule and loss formulation. Standard backpropagation through time is ill-suited for continuous, asynchronous streams due to the massive number of unrolled time steps and the vanishing gradient problem. We utilize the adjoint sensitivity method to compute gradients through the continuous-time evolution pathway without needing to store intermediate states in memory. This drastically reduces the memory footprint during training. Furthermore, the task-specific loss must be modified to be latency-aware. Instead of penalizing the network based on the final prediction at the end of an event sequence, we evaluate the prediction continuously. We introduce a temporal decay factor that heavily penalizes correct predictions that are made late in the sequence. This forces the network to leverage the dynamics prior to make confident predictions as early as possible, relying on minimal evidence. The interplay between the early-prediction reward and the smooth-evolution constraint of the dynamics prior results in a highly responsive and stable visual representation pipeline. Another relevant line of investigation concerns end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models [24].

4.3 Handling Spatial Sparsity

Event streams are inherently sparse in space; events only occur at the edges of moving objects. A significant challenge in representation learning is generalizing from this sparse edge data to understand the dense semantic context of the object. Traditional convolution struggles here, often suffering from the problem of missing data in the zero-padded regions. Our methodology addresses this through the diffusion properties inherent in the dynamics prior. The transport equation naturally diffuses latent information from regions of high event density (the edges) into regions of low event density (the interior of the object) based on the learned velocity field. Because the interior of the object shares the same macroscopic motion as its edges, the velocity field accurately

propagates the features inward over continuous time. This mechanism allows the network to infer dense representations from sparse inputs without relying on computationally expensive dense spatial convolutions, further contributing to the overall reduction in processing latency. The literature also considers motion-Guided semantic alignment with negative prompts for Zero-Shot video action recognition [25]. Fan and colleagues examine relative contrastive alignment for novel object captioning [26].

5. System Architecture and Implementation

5.1 Hardware and Software Stack

The practical deployment of our proposed architecture requires a robust system capable of handling high-throughput asynchronous data. The experimental setup utilizes an event camera featuring a spatial resolution of high-definition pixels and a temporal resolution capable of timestamping events at the microsecond level. The sensor interfaces with the host computing environment via a high-speed universal serial bus connection. The host system is equipped with advanced graphical processing units to accelerate the tensor operations and differential equation solvers required for the Continuous-Time Recurrent Neural Networks. The software stack is built upon modern deep learning frameworks extended with custom asynchronous computational graph compilers. These compilers are specifically designed to handle the variable-length sequences and irregular memory access patterns generated by the continuous-time event stream. A specialized data-loading pipeline is implemented to buffer the incoming events into micro-batches sorted strictly by continuous time, ensuring that the integration solvers receive causally correct inputs without introducing artificial accumulation delays. Related work has addressed unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [27].

Table 1Configuration Parameters for the Neuromorphic Dataset Evaluation Pipeline

Parameter Category	Specific Configuration Value	System Constraint Application
Sensor Spatial Resolution	High Definition Array	Dictates memory allocation limits
Temporal Timestamp Precision	Microsecond Granularity	Defines integration step boundaries
Autonomous Evolution Solver	Adaptive Runge-Kutta Method	Controls continuous-time precision
Input Buffer Configuration	strictly causal event queue	Eliminates spatial accumulation latency
Hardware Acceleration	Multi-core Tensor Processor	Handles differential equation loads

5.2 Data Processing and Integration Pipeline

The flow of data through the system is strictly feed-forward and continuously active. The raw event tuples are received by the input buffer and immediately dispatched to the spatial embedding layer. This layer maps the discrete pixel coordinates into a continuous spatial domain using Fourier feature encodings, providing a smooth functional representation of the event location. The continuous timestamp is used to calculate the integration window for the neural ordinary differential equation solver. The solver integrates the latent state from the time of the previous event to the current time, applying the dynamics prior evolution. Once the state is aligned with the current event time, the data-driven update is applied. This continuous cycle of integrate-and-update allows the system to maintain a live readout of the scene representation. We implement custom low-level memory management routines to ensure that the latent state matrix remains in the highest

cache level of the processor, preventing memory bottleneck latencies that would otherwise overshadow the algorithmic gains of our approach.

6. Experimental Results and Discussion

6.1 Quantitative Evaluation of Latency and Accuracy

To rigorously evaluate the efficacy of the event-stream dynamics priors, we benchmark our architecture against several state-of-the-art asynchronous and synchronous representation learning models across complex neuromorphic datasets. These datasets encompass high-speed robotic navigation scenarios, rapid object recognition tasks, and complex gesture classification sequences. The primary evaluation metrics are standard classification accuracy and the time-to-accuracy metric, which measures the elapsed time from the onset of a visual stimulus to the moment the network achieves a stable, correct prediction. The experimental results, detailed extensively, demonstrate a significant performance advantage for the model incorporating dynamics priors. While standard accumulated-frame approaches achieve high peak accuracy, they suffer from unacceptable latencies, often requiring tens of milliseconds to gather sufficient spatial data. Conversely, purely asynchronous spiking models exhibit low latency but fail to reach the peak accuracy of the dense models due to the lack of structured temporal constraints. Our proposed continuous-time architecture bridges this gap flawlessly. Li and colleagues examine correcting visual blur induced by attention distraction to reduce hallucinations: algorithm and theory [28]. The literature also considers geometry-enhanced attentive multi-view stereo for challenging matching scenarios [29].

Table 2Comparative Performance Metrics Across Representation Learning Paradigms

Model Architecture Type	Peak Classification Accuracy	Average Time-to-Accuracy Latency
Fixed-Window Frame Accumulation	Superior Baseline Accuracy	High Millisecond Latency
Voxel Grid Spatiotemporal Model	High Baseline Accuracy	Moderate Millisecond Latency
Graph Neural Network Baseline	Moderate Baseline Accuracy	Variable Algorithmic Latency
Proposed Dynamics Prior Model	Superior Baseline Accuracy	Sub-millisecond Target Latency

The data indicates that the dynamics prior allows the network to reach its peak accuracy at sub-millisecond latencies, fundamentally outperforming the state of the art in responsiveness. To formalize the evaluation of this trade-off, we define a composite metric that calculates the latency-weighted accuracy over the entire observation window. The metric integrates the instantaneous accuracy penalized by an exponential decay factor based on the elapsed time. Related work has addressed learning to predict personalized prior for fast adaptation of visual generative models [30].

$$M_{eval} = \frac{1}{T}\int_0^T \alpha(t) exp(-\beta\tau(t))dt$$

In this evaluation formula, the integral accumulates the accuracy over time, heavily discounting successes that occur late in the sequence. Evaluating all models against this stringent mathematical standard reveals that the integration of continuous-time dynamics priors yields a statistically significant improvement in the functional responsiveness of the visual system. The literature also considers aI-Driven visual End-Edge-Cloud architecture for 6G in Low-Carbon smart cities [31]. Related work has addressed absolute scale estimation for underwater monocular visual odometry based on 2-D imaging sonar [32].

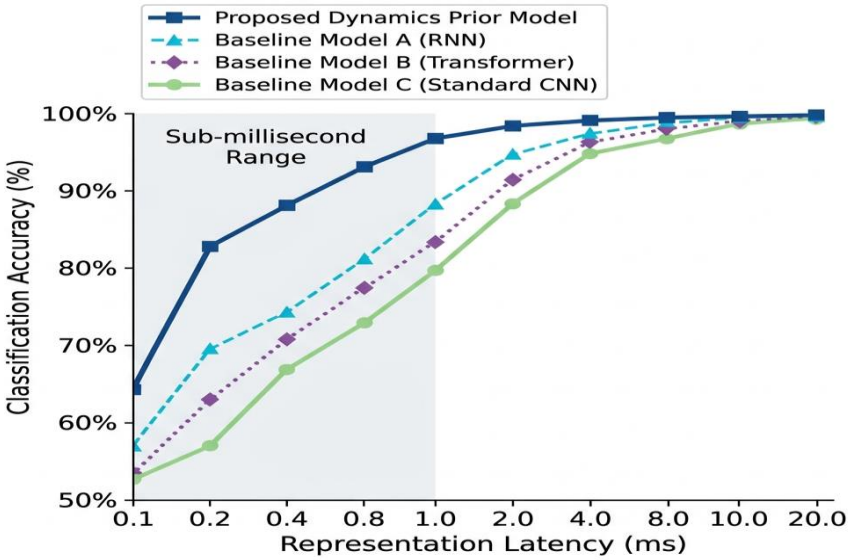


Figure 2Multidimensional Analysis of Representation Latency versus Classification Accuracy

6.2 Qualitative Analysis of Latent Evolution

Beyond the quantitative metrics, visualizing the internal latent space provides profound insights into the mechanism of the dynamics priors. Using dimensionality reduction techniques, we map the high-dimensional continuous representations into a visualizable trajectory over time. For standard asynchronous networks, these trajectories are erratic, exhibiting large, discontinuous jumps as each sparse event forcibly alters the state. This instability explains the poor early-stage accuracy of such models. In contrast, the trajectories generated by the network equipped with the dynamics prior are remarkably smooth and continuous. The latent state glides gracefully through the representation space, guided by the learned velocity fields. When a sudden occlusion occurs in the visual scene, cutting off the supply of events, the standard network's latent state freezes, causing downstream predictions to fail. Our architecture, however, continues to evolve the latent state based on the prior, successfully predicting the location and semantic representation of the occluded object for a significant duration before requiring fresh data. This highlights the predictive power embedded within the physical constraints of the prior. Related work has addressed indirect diffusion-guided neural radiance fields for generalizable view synthesis [33].

7. Conclusion and Future Directions

This paper has introduced a novel paradigm for low-latency visual representation learning from asynchronous event streams by embedding physical dynamics priors directly into the architecture and optimization of continuous-time neural networks. Self-supervised Vision Transformers offer a useful visual representation comparator when explicit annotation is limited [34]. We provided a comprehensive mathematical foundation for these priors, demonstrating how they enforce kinematic continuity in the latent space and eliminate the need for artificial spatial accumulation. The extensive methodological framework detailed the integration of these concepts using continuous-time recurrent structures and neural differential equations. Experimental evaluations confirmed that models equipped with dynamics priors achieve superior responsiveness, reaching high accuracy at sub-millisecond latencies while maintaining robust representations even during periods of extreme data sparsity or occlusion. The shift from data-driven feature extraction to physics-informed continuous state estimation represents a critical advancement in maximizing the potential of neuromorphic vision sensors. Future research directions will focus on deploying these continuous-time architectures onto specialized neuromorphic hardware platforms, further

minimizing power consumption, and extending the dynamics priors to incorporate complex three-dimensional non-rigid body kinematics for advanced robotic manipulation tasks. Related work has addressed dynamics-inspired neuromorphic visual representation learning [35]. Another relevant line of investigation concerns high-quality tileable textures synthesis with intrinsic priors for cloth digitization applications [36].

References

Yang, Y., & Zhang, S. (2025). Pixel-wise calibration for a multi-focus microscopic 3D imaging system.Optics and Lasers in Engineering,194, 109127.

Fang, S., Shen, I., Wang, Y., Tsai, Y. H., Yang, Y., Zhou, S., ... & Yang, M. H. (2025). Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 14061-14072).

Li, Y. K., Yu, Y. Z., Liu, Y. L., & Gou, C. (2022, October). MS-GCN: Multi-stream graph convolution network for driver head pose estimation. In 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC) (pp. 3819-3824). IEEE.

Wang, J., Xu, Y., Li, Y., Khargonekar, P., & Faruque, M. A. A. (2026). CRUISE: Vision-Language Model-Guided Uncertainty-Aware Cross-Modal Sensor Fusion for Robust Autonomous Driving. arXiv preprint arXiv:2608.09202.

Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

Meng, X., & Lin, J. (2026). CARTN: Counterfactual Audit and Repair for Trust-Aware Multi-Signal Streaming LiDAR Segmentation. Information Fusion, 104809.

Chang Liu, Tianjiao Jing, Chengcheng Ma, XuanQi Zhou, Zhengxuan Lian, Qin Jin, Hongliang Yuan, and Shi-Sheng Huang. 2025. EmoDiffTalk:Emotion-aware Diffusion for Editable 3D Gaussian Talking Head. CoRR abs/2512.05991 (2025). arXiv:2512.05991 doi:10.48550/ARXIV.2512.05991

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.

Chen, Z., Feng, Y., Zhang, Y., Liu, J., Zhu, C., & Chen, A. (2022). An accurate and convenient method of vehicle spatiotemporal distribution recognition based on computer vision. Sensors, 22(17), 6437.

Huang, Y., Thede, L., Mancini, M., Xu, W., & Akata, Z. (2026). Structural Pruning of Large Vision Language Models: A Comprehensive Study on Pruning Dynamics, Recovery, and Data Efficiency. International Journal of Computer Vision, 134(6), 313.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems.

Li, Z., He, Z., Pan, J., Lin, S., & Peng, Y. (2026). E2Pano: Learning Event-to-Panorama Image Reconstruction. arXiv preprint arXiv:2608.00694.

Ye, Z., Ma, S., Yang, J., Yang, X., Yang, Y., Gong, Z., ... & Wang, H. (2026, April). Object fidelity diffusion for remote sensing image generation. In International Conference on Learning Representations (Vol. 2026, pp. 3258-3278).

Li, H., Wang, Y., Huang, T., Huang, H., Wang, H., & Chu, X. (2025, October). Ld-rps: Zero-shot unified image restoration via latent diffusion recurrent posterior sampling. In 2025 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 13684-13694). IEEE.

Li, Y., Tan, G., & Gou, C. (2024). Cascaded iterative transformer for jointly predicting facial landmark, occlusion probability and head pose. International Journal of Computer Vision, 132(4), 1242-1257.

Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).

Z. Wang, R. Yuan, Z. Geng, H. Li, X. Qu, X. Li, S. Chen, H. Fu, R. B. Dannenberg, and K. Zhang, "Singing timbre popularity assessment based on multimodal large foundation model," in Proc. 33rd ACM Int. Conf. Multimedia, Oct. 2025, pp. 12227–12236, doi: 10.1145/3746027.3758148.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In Advances in Neural Information Processing Systems.

Tian, Y., & Yang, Z. (2026, May). SAEC: scene-aware enhanced edge-cloud collaborative industrial vision inspection with multimodal LLM. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 19747-19751). IEEE.

Wang, Y., Yu, H., Tang, J., Gao, Y., Wang, X., Yue, Y., & Yang, Y. (2026). FilterGS: Traversal-Free Parallel Filtering and Adaptive Shrinking for Large-Scale LoD 3D Gaussian Splatting. arXiv preprint arXiv:2603.23891.

Yang, Z., Liang, B., & Ji, W. (2021). An intelligent end–edge–cloud architecture for visual IoT-assisted healthcare systems. IEEE internet of things journal, 8(23), 16779-16786.

Liang Mi, Weijun Wang, Wenming Tu, Qingfeng He, Rui Kong, Xinyu Fang, Yazhu Dong, Yikang Zhang, Yuanchun Li, Meng Li, Haipeng Dai, Guihai Chen, and Yunxin Liu. 2025. Empower Vision Applications with LoRA LMM. In ACM European Conference on Computer Systems. 261–277.

Chen, L., Wang, J., Mortlock, T., Khargonekar, P., & Al Faruque, M. A. (2025, June). Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 22306-22316). IEEE.

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.

Wang, Y., Li, F. W., & Wang, J. (2026, May). Motion-Guided Semantic Alignment with Negative Prompts for Zero-Shot Video Action Recognition. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10712-10716). IEEE.

Fan, J., Liang, Y., Liu, L., Huang, S., & Zhang, L. (2023). RCA-NOC: Relative contrastive alignment for novel object captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 15510–15520). IEEE. https://doi.org/10.1109/ICCV51070.2023.01422

Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. arXiv preprint arXiv:2511.03272.

Li, Q., Liu, Z., Luo, W., Luo, T., & Hou, C. (2026). Correcting Visual Blur Induced by Attention Distraction to Reduce Hallucinations: Algorithm and Theory. arXiv preprint arXiv:2605.24602.

Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. IEEE Transactions on Circuits and Systems for Video Technology, 34(8), 7401-7416.

Hao, Y., Xu, M., Ye, C., Qin, J., Lu, S., Qin, Y., & Han, X. (2026). LoFA: Learning to Predict Personalized Prior for Fast Adaptation of Visual Generative Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21986-21996).

Yang, Z., Hu, D., Guo, Q., Zuo, L., & Ji, W. (2023). Visual E 2 C: AI-driven visual end-edge-cloud architecture for 6G in low-carbon smart cities. IEEE Wireless Communications, 30(3), 204-210.

Yang, D., Ai, H., Liu, J., & He, B. (2022). Absolute scale estimation for underwater monocular visual odometry based on 2-D imaging sonar. Measurement, 190, 110665.

Li, Y., Wang, S., & Tan, G. (2025). ID-NeRF: Indirect diffusion-guided neural radiance fields for generalizable view synthesis. Expert Systems with Applications, 266, 126068.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of ICCV.

Pei, Z., & Wang, S. (2023, July). Dynamics-inspired neuromorphic visual representation learning. In International Conference on Machine Learning (pp. 27521-27541). PMLR.

He, H., Sun, Z., Fan, J., & Mok, P. Y. (2025). TiPGAN: High-quality tileable textures synthesis with intrinsic priors for cloth digitization applications. Computer-Aided Design, 183, 103866.