FIDELITY-AWARE EVENT REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Event cameras excel in dynamic sensing. Harnessing this potential hinges on constructing faithful, compact, and statistically stable representations from irregular events, a fundamental yet nontrivial challenge. Existing event representations face a persistent dilemma: information fidelity is traded for statistical tractability through progressively stronger structural constraints. To reconcile this tension, we propose a fidelity-aware event representation framework that distills asynchronous events into physics-grounded latents while preserving temporally structured event information. Raw events are first structured into conflict-free volumes using the coarsest temporal discretization observed to avoid intra-bin event collisions. We then employ a forward-physics transformation that integrates sparse differential signals into relative log-intensity trajectories. This transformation is exactly invertible by temporal differencing, reorganizing the information retained after discretization into a more coherent temporal form without further information loss, thereby bridging information fidelity and statistical tractability. A variational autoencoder subsequently maps these proxy trajectories into compact and distributionally regular latents that serve as model-ready representations. Extensive experiments demonstrate faithful and stable event latents that transfer across diverse domains and visual objectives, spanning semantic understanding, motion perception, and photometric reconstruction. These results support physics-guided latent learning as an effective approach to event representation. *Code will be publicly available.*

1 INTRODUCTION

Transcending frame-based acquisition, event cameras operate as neuromorphic observers (Lichtsteiner et al., 2006; Posch et al., 2010; Delbrück et al., 2010) that free visual sensing from fixed temporal intervals. By asynchronously reporting per-pixel intensity changes, they deliver microsecond latency, extreme dynamic range, and high energy efficiency. These properties have enabled a broad spectrum of event-based vision applications (Lin et al., 2025; Yu et al., 2024; Xie et al., 2025; Xu et al., 2026). Yet a fundamental gap remains between irregular event streams and the structured inputs favored by modern learning architectures. Bridging this gap requires a principled event representation that renders asynchronous spatio-temporal streams compact and statistically tractable while preserving their temporally structured information.

Since the advent of event cameras, diverse representations have emerged to bridge event signals and learning-based vision systems. As illustrated in Fig. 1, early approaches represent events as point-wise sets (Bi et al., 2020; Li et al., 2021; Chen et al., 2022), retaining individual events while remaining highly irregular. Pixel-level sequence representations introduce temporal organization at each spatial location (Sironi et al., 2018; Ramesh et al., 2019; Baldwin et al., 2022), but often summarize recent activity or maintain truncated temporal histories. Grid-based tensors impose stronger spatio-temporal regularity through temporal aggregation (Maqueda et al., 2018; Zhu et al., 2019; Nam et al., 2022), providing convenient interfaces to conventional vision backbones but potentially merging multiple co-located events within discrete temporal intervals. This progression exposes a persistent dilemma: statistical tractability is often achieved through temporal truncation, summarization, or aggregation that irreversibly discards fine-grained temporal information during representation construction.

This dilemma shifts the key question from how much structure to impose to how to structure events without unnecessary information loss. An individual event provides only an isolated thresholded differential measurement of log intensity; its physical meaning emerges from the continuous temporal evolution of the imaging process. Individually, such measurements are sparse and fragmented,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

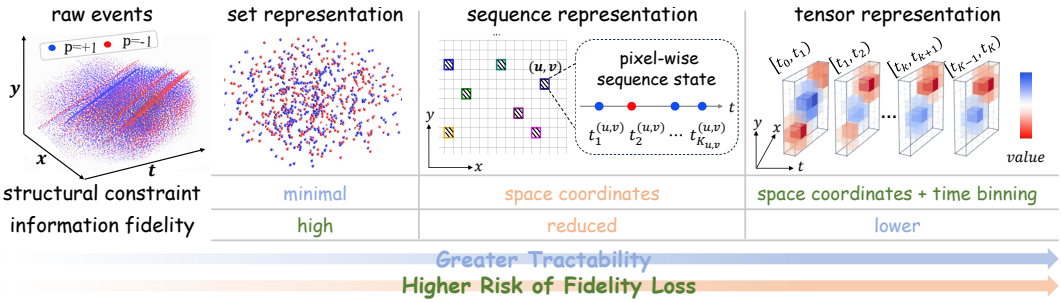


Figure 1: **Event representation dilemma**. From raw sets to spatially constrained sequences and spatio-temporally discretized tensors, increasingly structured representations tend to improve statistical tractability for learning, while temporal truncation and aggregation can irreversibly erode fine-grained temporal information.

whereas their collective evolution reflects the underlying intensity dynamics encoded by the sensor. This observation suggests that statistical tractability need not come at the cost of fine-grained temporal information. Instead, event-generation physics provides a natural prior for reorganizing irregular measurements into a more coherent form while retaining their encoded information. We therefore pursue statistical tractability by reorganizing event information before compressing its content.

Motivated by this principle, we propose a **fidelity-aware event representation framework** that distills asynchronous events into physics-grounded latent representations. Our framework first organizes raw events into conflict-free volumes using the coarsest temporal discretization observed to avoid intra-bin event collisions, while limiting unnecessary sparsity. This preserves the separation of co-located events under temporal discretization rather than prematurely merging them through coarse aggregation. We then employ a *forward-physics transformation* that integrates the sparse differential signals over time into relative log-intensity trajectories. Crucially, this transformation is exactly invertible by temporal differencing, reorganizing the information retained after conflict-free voxelization into a denser, more coherent, and temporally stable form without further information loss. Building on this physics-grounded proxy, a variational autoencoder (Kingma & Welling, 2013; Diederik & Max, 2019) compresses the structured trajectories into compact and distributionally regular latent variables. Rather than relying on substantial architectural customization, the resulting latents provide model-ready representations that encode reusable event information across scenes and visual tasks.

We evaluate the resulting representation from two complementary perspectives: its intrinsic representation quality and its utility for downstream vision. First, we jointly examine fidelity and cross-domain transferability by assessing whether the physics-grounded proxy can be faithfully reconstructed from the compressed latent and whether the learned latent representations exhibit stable statistics across diverse scenes, event cameras, motion conditions, and data sources. Second, we compare our latent representation with existing event representations across diverse downstream visual tasks to assess its utility for different visual objectives. Our experiments span *semantic understanding*, *motion perception*, and *photometric reconstruction*, providing complementary evaluations from high-level recognition to low-level signal recovery. Together, these results support physics-guided latent learning as an effective approach to event representation.

In summary, the main contributions of this paper are:

- We revisit the persistent fidelity–tractability dilemma in event representation and identify physics-guided reorganization as a means to improve statistical structure without unnecessary information loss;
- We propose a physics-grounded latent representation framework that combines conflict-free volumization with an exactly invertible forward-physics transformation. By reorganizing sparse differential events into coherent relative log-intensity trajectories before latent compression, the framework improves statistical tractability without further information loss beyond temporal discretization;
- We compress the resulting physics-grounded proxy into compact and distributionally regular event latents, and systematically evaluate their fidelity, cross-domain transferability, and downstream utility across semantic understanding, motion perception, and photometric reconstruction.

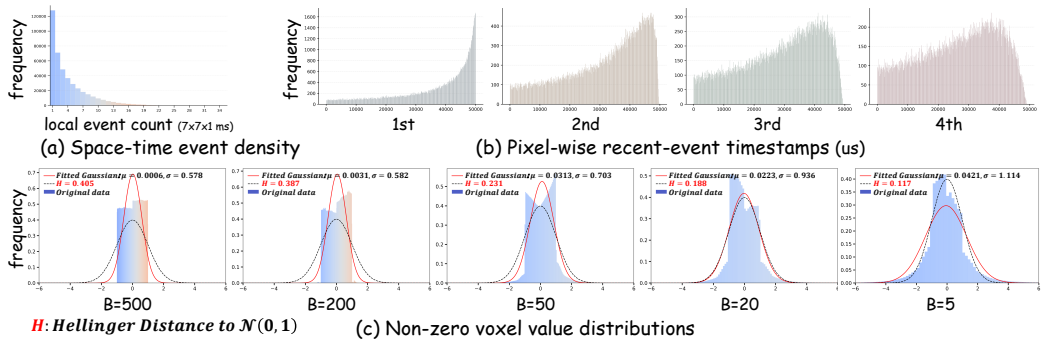


Figure 2: **Event representation statistics** (DSEC (Gehrig et al., 2021a), 50ms). (a) Raw events exhibit highly non-uniform local densities. (b) Pixel-wise recent-event timestamps show strongly varying temporal statistics. (c) Time binning progressively regularizes voxel statistics while merging intra-bin temporal details. These observations highlight the tension between statistical tractability and temporal fidelity.

2 Related Work

Event representations bridge irregular event signals and the input formats used by modern learning systems. Existing paradigms span different levels of spatio-temporal structure, from point-wise sets and pixel-level sequences to grid-based tensors with more regular statistics. Across these paradigms, a central challenge is to improve statistical tractability without unnecessarily sacrificing fine-grained temporal information. See *Appendix A.1* for a more detailed analysis.

Point-wise Set Representations. Set-based paradigms retain individual events as point-wise sets. Early approaches (Wang et al., 2019b; Yang et al., 2019) process event sets with point-based networks, while later methods employ graph-based processing (Bi et al., 2020; Li et al., 2021; Deng et al., 2022) or hybrid/state-space backbones (Ren et al., 2023; 2025). Retaining individual events preserves fine-grained temporal information, but also leads to varying set cardinality and non-uniform local density (Figure 2(a)), complicating event sampling, scalable processing, and robust feature learning for complex vision tasks.

Pixel-level Sequence Representations. Pixel-level representations organize events temporally at each spatial location. Time surfaces summarize recent activity (Lagorce et al., 2016; Sironi et al., 2018; Ramesh et al., 2019; Kim et al., 2021), while other methods maintain per-pixel states (Cannici et al., 2020; Annamalai et al., 2022; Baldwin et al., 2022) and can be paired with state-space or spiking architectures (Zubic et al., 2024; Zhu et al., 2022a; Mészáros et al., 2025). These representations preserve fine-grained local timing but typically retain only truncated or recent-event histories, with temporal statistics varying substantially across pixels (Figure 2(b)).

Grid-based Tensor Representations. Grid-based tensorization has become the predominant interface between event streams and modern vision backbones. Event Frames (Rebecq et al., 2017) and Histograms (Maqueda et al., 2018) first aggregate event polarities or counts on spatial grids, while Voxel Grids (Zhu et al., 2019) extend this construction along time through temporal binning. Subsequent representations increase information capacity or adapt the aggregation scheme through higher-dimensional stacking, mixed-density structures, or adaptive temporal organization (Teng et al., 2022; Nam et al., 2022; Liu et al., 2023a). Other works improve tensor quality through motion compensation or timestamp correction (Chen et al., 2025c; Liu et al., 2025b), combine complementary descriptors (Wang et al., 2019a; Fan et al., 2025), or optimize representational configurations directly (Zubić et al., 2023). Their regular tensor structure enables seamless integration with CNNs, transformers, and pretrained vision models, driving their widespread adoption in event-based vision systems (Chen et al., 2024; Kong et al., 2024; Zhu et al., 2025; Chen et al., 2026; 2025b; Yang et al., 2025).

Grid-based tensor representations typically improve tractability through temporal aggregation, at the cost of irreversible loss of fine-grained temporal information (Figure 2(c)); see *Appendix A.1.3* for details. In contrast, our framework first constructs a conflict-free discretization to avoid observable intra-bin event collisions, then applies an exactly invertible forward-physics transformation to reorganize the retained event information before learning a compact latent representation.

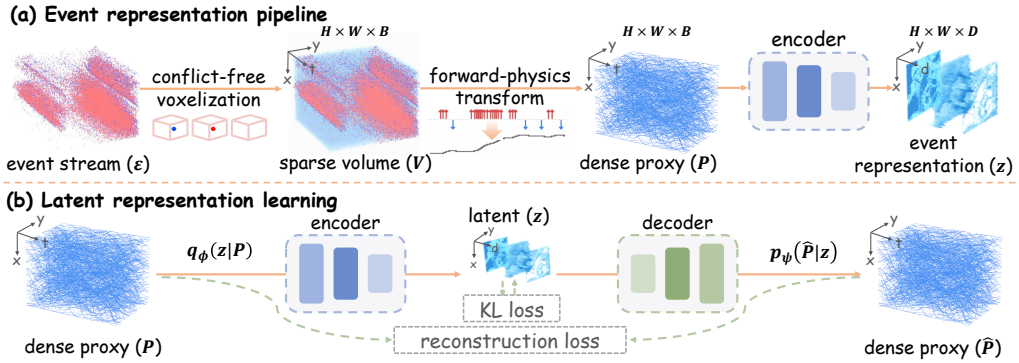


Figure 3: **Overview of the proposed framework**. The framework learns compact and statistically regular event latents through fidelity-aware structuring and physics-guided reorganization.

3 Proposed Method

An event camera outputs an asynchronous event stream $\mathcal{E} = \{e_i\}_{i=1}^{N}$, where $e_i = (x_i, y_i, t_i, p_i)$ denotes pixel location $(x_i, y_i) \in [\![0, W-1]\!] \times [\![0, H-1]\!]$, timestamp t_i, and polarity $p_i \in \{-1, +1\}$. Let $L = \log I$. An event is triggered when $L(x_i, y_i, t_i) - L(x_i, y_i, t_i^-) = p_i c$, where t_i^- is the preceding reference time at the same pixel. For a temporal window $[t_s, t_e]$ of duration $\Delta t = t_e - t_s$, we encode $\mathcal{E}_{[t_s,t_e]}$ into a compact event latent $\mathbf{z} \in \mathbb{R}^{H\times W\times D}$. As shown in Fig. 3, our framework consists of *conflict-free voxelization* for collision-free event structuring (Sec. 3.1), a *forward-physics transformation* that reorganizes the sparse event volume into a structured physics-grounded proxy (Sec. 3.2), and a *variational autoencoder* that learns a compact and distributionally regular latent representation (Sec. 3.3).

3.1 Conflict-free Event Voxelization

Conventional event voxelization discretizes a temporal window of duration Δt with bin width δ_t, yielding $B = \lceil \Delta t/\delta_t \rceil$ bins. Coarse bins cause *time aliasing*, merging co-located events and losing their temporal ordering, whereas overly fine bins increase sparsity and hinder representation learning. We therefore seek the coarsest bin width that avoids observable intra-bin event collisions.

Conflict-free Condition. Given δ_t, let $\tau_b = t_s + b\delta_t$ and $n_b(x, y) = \#\{\, i \mid (x_i, y_i) = (x, y),\ t_i \in [\tau_b, \tau_{b+1})\,\}$ be the event count in cell (x, y, b). The voxelization is *conflict-free* iff $\max_{x,y,b} n_b(x, y) \le 1$. Under this condition, $\mathbf{V} \in \mathbb{R}^{H\times W\times B}$ with $\mathbf{V}(x, y, b) = \sum_{i:(x_i,y_i)=(x,y),\, t_i\in[\tau_b,\tau_{b+1})} p_i$ yields a collision-free event assignment under temporal discretization. This preserves the separation and polarity of events at the chosen temporal resolution, while exact sub-bin timestamps are not retained.

Empirical Temporal Anchor. Among conflict-free voxelizations, a larger δ_t yields fewer bins and hence a more compact representation. We therefore define the coarsest admissible bin width as

$$\delta_t^\star = \max\left\{\delta_t > 0 \;\middle|\; \max_{x,y,b} n_b(x, y) \le 1\right\}. \tag{1}$$

Estimating $\delta_t^\star$ from the minimum per-pixel inter-event gap is fragile, as isolated near-zero intervals can dominate. We therefore sweep δ_t and select the coarsest bin width at which event restoration saturates. As shown in Fig. 4, restoration reaches 100% at $\delta_t = 100\,\mu$s across the tested datasets, sensors, and 10–100 ms windows, with no observable intra-bin event collisions; finer discretization only increases sparsity. We thus set $\delta_t^\star = 100\,\mu$s as an *empirical temporal anchor*, with $B^\star = \lceil \Delta t/\delta_t^\star \rceil$. Notably, this scale falls within the range of reported pixel-response latencies, including 100–220 μs for the Sony–Prophesee IMX636 (Guo et al., 2024) and approximately 100 μs for the DVS240 under the

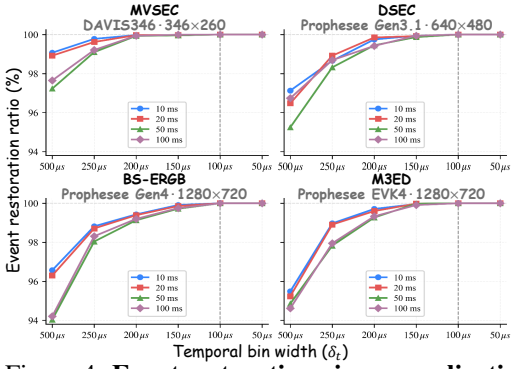


Figure 4: **Event restoration via re-voxelization across temporal bin widths.** Restoration ratio denotes the recovered-event fraction.

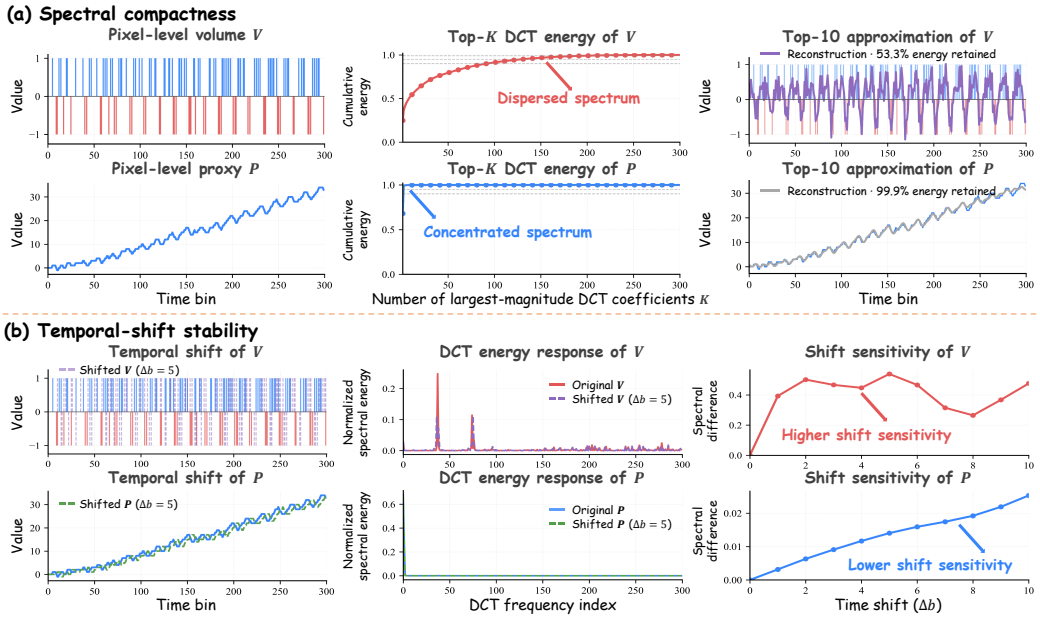


Figure 5: **Spectral compactness and shift stability of the physics-grounded proxy.** (a) Compared with the sparse volume $\mathbf{V}$, the proxy $\mathbf{P}$ concentrates spectral energy into fewer DCT coefficients, revealing greater intrinsic temporal compressibility. (b) Under controlled temporal shifts, $\mathbf{P}$ exhibits smaller changes in its natural-order DCT energy response, indicating greater shift stability.

tested conditions (Holešovskỳ et al., 2021), providing a physical interpretation of the empirical scale without implying a universal collision-free bound.

3.2 Forward-physics Volume Regularization

Under the conflict-free condition in Sec. 3.1, voxelization produces a sparse volume without intra-bin event collisions. However, the resulting representation remains statistically unfavorable for learning: only approximately 0.1%–1.0% of the volume is occupied, and each pixel-wise temporal signal $\mathbf{V}(x, y, :)$ consists primarily of isolated bipolar impulses. As illustrated in Fig. 5, this sparse impulse structure results in poor spectral compactness and high sensitivity to small temporal shifts.

To improve these statistical properties without compromising the information encoded in $\mathbf{V}$, we exploit the event-generation physics to reorganize it into a more structured form. Since an event encodes a thresholded differential observation of log intensity, temporal accumulation yields an event-encoded proxy of the underlying relative log-intensity variation. Specifically, for each pixel (x, y), we define

$$\mathbf{P}(x, y, b) = \sum_{b'=0}^{b} \mathbf{V}(x, y, b'), \qquad b = 0, \ldots, B^\star - 1, \tag{2}$$

where $\mathbf{P}$ serves as a piecewise-constant proxy for the relative log-intensity variation $\Delta \log I(x, y, \tau_b)$.

Importantly, the transformation changes the *organization* of event information rather than its *content*. The resulting proxy $\mathbf{P} \in \mathbb{R}^{H \times W \times B^\star}$ provides three favorable properties. **(i) Spectral Compactness.** Temporal accumulation converts isolated impulses into persistent temporal trajectories, concentrating the dispersed spectral energy of $\mathbf{V}$ into a substantially more compact form in $\mathbf{P}$. As shown in Fig. 5(a), DCT analysis reveals that this reorganization substantially improves intrinsic temporal compressibility, with spectral energy concentrated in a small number of dominant components. **(ii) Shift Stability.** Persistent temporal trajectories make $\mathbf{P}$ substantially less sensitive to small timing perturbations than $\mathbf{V}$, as demonstrated in Fig. 5(b). This stability also motivates using $\mathbf{P}$ for fidelity measurement, since minor temporal misalignments have a reduced influence on the measured discrepancy. **(iii) Information Preservation.** The transformation is exactly invertible by temporal differencing:

$$\mathbf{V}(x, y, b) = \mathbf{P}(x, y, b) - \mathbf{P}(x, y, b-1), \qquad b \geq 1, \tag{3}$$

with $\mathbf{V}(x, y, 0) = \mathbf{P}(x, y, 0)$. Thus, the mapping from $\mathbf{V}$ to $\mathbf{P}$ introduces no additional information loss beyond the preceding temporal discretization.

The forward-physics transformation introduces no learnable parameters or auxiliary supervision. By reorganizing sparse differential measurements into coherent relative log-intensity trajectories, it provides a more statistically tractable proxy for subsequent latent representation learning.

3.3 Latent Representation Learning

Variational Latent Encoding. Having reorganized irregular asynchronous events into a statistically structured physics-grounded proxy, we further distill $\mathbf{P}$ into a compact latent representation using a lightweight variational autoencoder ($\sim$2M parameters). We focus on the representation and training objective rather than substantial architectural customization. Architectural details are provided in *Appendix A.2.1*.

Since each entry of $\mathbf{P}$ accumulates at most $B^\star$ polarities, we normalize it as $\widetilde{\mathbf{P}} = \mathbf{P}/B^\star \in [-1,1]^{H\times W\times B^\star}$, providing a fixed input range for latent encoding. The encoder $q_\phi(\mathbf{z} \mid \widetilde{\mathbf{P}})$ parameterizes a diagonal Gaussian posterior,

$$q_\phi(\mathbf{z} \mid \widetilde{\mathbf{P}}) = \mathcal{N}\Big(\mathbf{z}; \boldsymbol{\mu}_\phi(\widetilde{\mathbf{P}}), \operatorname{diag}\Big(\boldsymbol{\sigma}_\phi^2(\widetilde{\mathbf{P}})\Big)\Big)\,, \tag{4}$$

from which latent samples are obtained via $\mathbf{z} = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The decoder reconstructs the normalized proxy $\hat{\widetilde{\mathbf{P}}}$ from the latent code $\mathbf{z}$.

Event-aware Latent Objective. Although the physics-grounded proxy is substantially denser than the original event volume, $\widetilde{\mathbf{P}}$ still contains many zero-valued entries. A uniformly weighted reconstruction loss would therefore be dominated by these entries and underemphasize active proxy regions. We identify $\mathbf{M}(x,y,b) = \mathbf{1}[\mathbf{P}(x,y,b) \neq 0]$ and restrict reconstruction supervision to these locations. Together with KL regularization of the latent distribution, the overall objective is

$$\mathcal{L}_{\text{latent}} = \Big\| \mathbf{M} \odot \Big(\hat{\widetilde{\mathbf{P}}} - \widetilde{\mathbf{P}}\Big) \Big\|_1 + \lambda_{\text{kl}} \mathrm{KL}\Big(q_\phi(\mathbf{z} \mid \widetilde{\mathbf{P}}) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\Big)\,, \tag{5}$$

where λ_{kl} controls the strength of latent KL regularization. Combined with the physics-grounded proxy, this objective learns compact and distributionally regular event representations.

4 Experiments

4.1 Evaluation of Latent Learning

Training Setup. We train the variational autoencoder from scratch on the DSEC training split (Gehrig et al., 2021a). Event segments are randomly sampled along continuous streams and converted into normalized proxies $\widetilde{\mathbf{P}} \in \mathbb{R}^{H\times W\times B}$, where $B = \lceil \Delta t/\delta_t^\star \rceil$ and $\delta_t^\star = 100\,\mu$s. Random sampling provides a large pool of segments with diverse temporal content. We denote a model configuration by **`T-B-D`**, where **`T`** is the window duration in milliseconds, **`B`** the corresponding temporal-bin count, and **`D`** the latent channel dimension: $\widetilde{\mathbf{P}} \in \mathbb{R}^{H\times W\times B} \to \mathbf{z} \in \mathbb{R}^{H\times W\times D}$. For example, **`T30-B300-D3`** denotes a 30 ms window, 300 bins, and a 3-channel latent.

Evaluation and Metrics. We evaluate on the DSEC test split and seven cross-domain datasets spanning diverse scenes, event cameras, motion conditions, and data sources: DDD17 (Binas et al., 2017), MVSEC (Zhu et al., 2018), M3ED (Chaney et al., 2023), RLED (Liu et al., 2024), ERFX170 (Kim et al., 2023), and two VID2E-synthesized datasets (Gehrig et al., 2020), GoPro (Nah et al., 2019) and SDSD (Wang et al., 2021c), without dataset-specific adaptation. We vary the event window Δt (10–100 ms), latent dimension $D \in \{1,3,5,8,10\}$, and KL weight $\lambda_{\text{kl}} \in \{10^{-5},\ldots,10^{-1}\}$ to study temporal scale, compression level, and latent regularization, respectively. Latent quality is evaluated by **PSNR-active**, **ST-SSIM**, and **KL/dim**, measuring reconstruction fidelity on event-active trajectories, local spatio-temporal structure preservation, and latent regularity, respectively. Together, these metrics characterize proxy reconstruction fidelity and latent regularity. Additional implementation details are provided in *Appendix A.2.2*.

Results of Representation Learning. Fig. 6 shows that the learned event latents preserve the physics-grounded proxy with high fidelity across in-domain and cross-domain settings. Despite training only on DSEC, the model achieves high PSNR-active and near-unity ST-SSIM on seven cross-domain datasets spanning diverse sensors, scenes, resolutions, motion conditions, and data sources, without

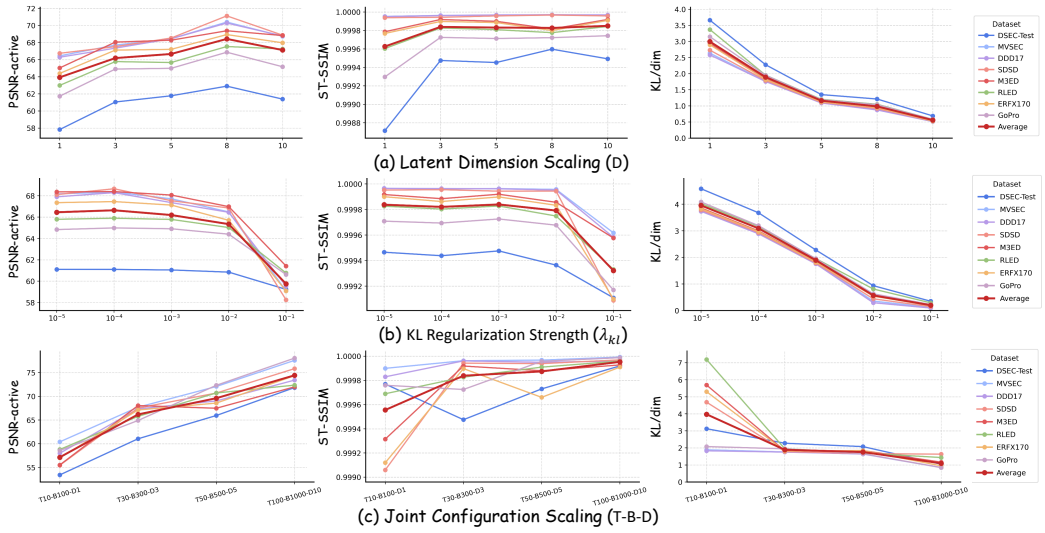


Figure 6: **Quantitative analysis of latent learning.** (a) Effect of latent dimension **D** under fixed **T30−B300**. (b) Effect of KL regularization under fixed **T30−B300−D3**. (c) Joint scaling of **T−B−D**. All settings are evaluated on DSEC and seven cross-domain datasets.

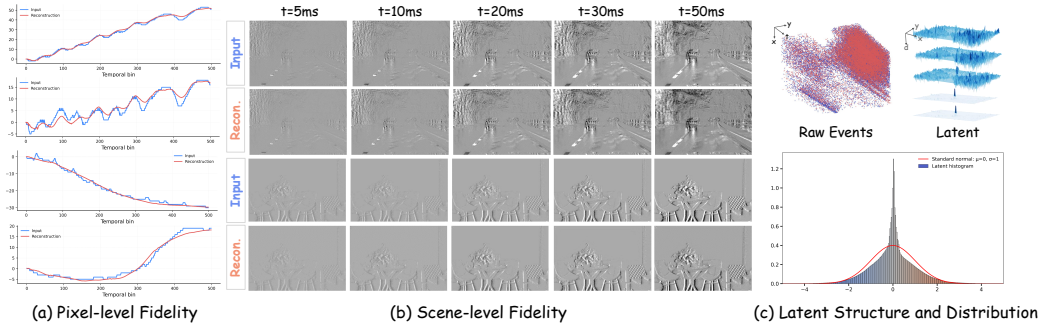


Figure 7: **Qualitative properties of the latent representation.** Shown with **T50−B500−D5**. (a) Reconstructed pixel-wise trajectories preserve temporal dynamics. (b) Scene-level reconstructions retain spatio-temporal structure across in-domain and cross-domain data. (c) Latent responses exhibit structured spatial patterns and regularized distributions.

dataset-specific adaptation. This consistency suggests that the latent captures transferable event structure beyond the training domain.

The three analyses reveal a clear capacity–regularization trade-off. Increasing the latent dimension improves reconstruction fidelity while reducing KL/dim, with gains saturating at larger dimensions. Increasing λ_{kl} strengthens latent regularization, whereas excessive regularization degrades reconstruction fidelity. Joint **T−B−D** scaling further improves reconstruction as temporal context, temporal support, and latent capacity increase together. These trends remain consistent across in-domain and cross-domain datasets.

4.2 Evaluation on Downstream Tasks

We further evaluate the learned event representation on three complementary dense visual tasks: semantic segmentation, optical flow estimation, and video reconstruction. Owing to its regular tensor form, the proposed latent can be directly integrated into existing event-based pipelines without task-specific redesign. We instantiate this controlled replacement in three voxel-grid-based pipelines: **ScaleEvent** for semantic segmentation (Chen et al., 2026), **STSC** for optical flow estimation (Hu et al., 2026), and **V2V** for video reconstruction (Lou et al., 2026). For each task, we retain the original architecture, training schedule, optimization, data processing, and evaluation protocol, replacing only the voxel-grid input with our event latent. This controlled replacement enables a matched comparison of input representations while keeping the downstream framework unchanged. Additional details are provided in *Appendix A.2.4*.

Table 1: **Semantic segmentation comparison.** Best **bold** and second-best underlined.

Method	Backbone	DDD17-Seg		DSEC-Sem	
		Acc ↑	mIoU ↑	Acc ↑	mIoU ↑
DTL Wang et al. (2021a)	ResNet-50	-	58.80	-	-
EVDistill Wang et al. (2021b)	ResNet-34	-	58.02	-	-
ESS-Sup Sun et al. (2022)	E2VID	91.08	61.37	89.37	53.29
EvSegformer Jia et al. (2023)	MiT-B1	**94.72**	54.41	-	-
ESEG Zhao et al. (2025)	MiT-B1	90.68	59.97	91.47	57.55
KWYAF Li et al. (2025)	MiT-B0	91.32	62.41	90.87	57.75
ECDDP Yang et al. (2024)	Swin-T/7	-	62.56	-	61.25
OpenESS Kong et al. (2024)	E2VID	91.05	63.00	90.21	57.21
STP Liang et al. (2025)	Swin-T/7	-	63.29	-	62.05
ScaleEvent (Chen et al., 2026)	ViT-L/16	92.62	65.08	93.10	69.65
Ours (ScaleEvent)	ViT-L/16	92.82	**65.95**	**93.64**	**71.84**

Table 2: **Optical flow estimation comparison.** Best **bold** and second-best underlined.

Method	EPE ↓	3PE ↓	2PE ↓	1PE ↓	AE ↓
E-RAFT (Gehrig et al., 2021b)	0.788	2.68	4.74	12.74	2.85
ADMFlow (Luo et al., 2023)	0.779	2.65	4.67	12.52	2.84
EEMFlow+ (Luo et al., 2024)	0.751	2.15	3.93	11.40	2.67
TMA (Liu et al., 2023b)	0.743	2.30	3.97	10.86	2.68
EDCFlow (Liu et al., 2025a)	0.720	2.10	3.60	10.00	2.6
IDNet (Wu et al., 2024)	0.719	2.04	3.50	10.07	2.72
ResFlow (Zhou et al., 2025)	0.754	2.50	4.24	11.22	2.73
BFlow (Gehrig et al., 2024)	0.750	2.44	4.41	11.90	2.68
STSC (Hu et al., 2026)	0.677	1.74	2.90	8.33	2.56
Ours (STSC)	**0.646**	**1.52**	**2.47**	**7.61**	**2.36**

Table 3: **Video reconstruction comparison**. **Voxel-Intp.** denotes interpolated voxel grids and is used by default unless otherwise specified, while **Voxel-Disc.** denotes the discrete voxel grids of SBT (Mostafavi et al., 2021). Methods marked with † provide public training code.

Method	Train Data	Representation	IJRR			MVSEC			HQF		
			MSE ↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS ↓	MSE ↓	SSIM ↑	LPIPS ↓
E2VID+† (Stoffregen et al., 2020)	ESIM-280	Voxel-Intp.	0.047	0.578	0.217	0.153	0.326	0.531	0.048	0.605	0.285
FireNet+† (Stoffregen et al., 2020)	ESIM-280	Voxel-Intp.	0.063	0.526	0.290	0.218	0.294	0.570	0.040	0.577	0.314
ET-Net (Weng et al., 2021)	ESIM-280	Voxel-Intp.	0.049	0.588	0.239	0.111	0.360	0.491	0.035	0.642	0.274
HyperE2VID (Ercan et al., 2024)	ESIM-280	Voxel-Intp.	0.036	0.623	0.231	0.071	0.418	0.473	0.032	0.646	0.265
V2V-E2VID† (Lou et al., 2026)	ESIM-280	Voxel-Intp.	0.070	0.561	0.265	0.150	0.338	0.509	0.043	0.597	0.304
V2V-E2VID† (Lou et al., 2026)	ESIM-280	Voxel-Disc.	0.068	0.564	0.264	0.147	0.342	0.503	0.041	0.604	0.297
Ours (V2V-E2VID)	ESIM-280	Latent	0.043	0.596	0.236	0.089	0.376	0.490	0.038	0.629	0.283

Semantic Segmentation. Following ScaleEvent (Chen et al., 2026), we evaluate semantic segmentation on DDD17-Seg (Alonso & Murillo, 2019) and DSEC-Semantic (Sun et al., 2022). Performance is measured by mean intersection over union (mIoU) and average class accuracy (Acc). We use `T50-B500-D3` for DDD17-Seg and `T30-B300-D3` for DSEC-Semantic.

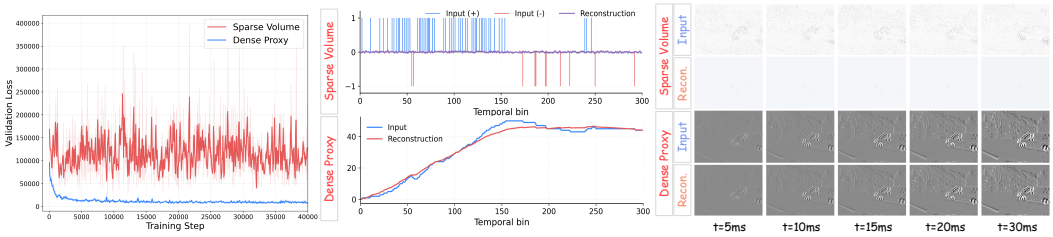
Table 1 shows that our representation achieves the best mIoU on both benchmarks, reaching 65.95% on DDD17-Seg and 71.84% on DSEC-Semantic. Under the same ViT-L/16 backbone, it improves over ScaleEvent by 0.87 and 2.19 mIoU points, respectively, while also yielding higher accuracy on both datasets. With the downstream architecture and training protocol unchanged, these gains indicate that the proposed latent retains task-relevant event information while providing a statistically tractable representation for dense semantic prediction.

Optical Flow Estimation. Following STSC (Hu et al., 2026), we evaluate optical flow estimation on DSEC-Flow (Gehrig et al., 2021a). The primary metric is End-Point Error (EPE), defined as the average Euclidean distance between predicted and ground-truth flow. We further report 3PE, 2PE, and 1PE, denoting the percentage of pixels whose EPE exceeds 3, 2, and 1 pixels, respectively, together with Angular Error (AE). We use `T10-B100-D1` for DSEC-Flow.

As shown in Table 2, our representation achieves the best results across all DSEC-Flow metrics, with an EPE of 0.646, 3PE/2PE/1PE of $1.52/2.47/7.61$, and an AE of 2.36. Compared with STSC (Hu et al., 2026), it reduces EPE from 0.677 to 0.646 while consistently improving all thresholded errors and AE. Under an unchanged downstream pipeline, these gains indicate that the proposed latent retains motion information useful for dense optical flow estimation.

Video Reconstruction Following V2V (Lou et al., 2026), we train the classic E2VID (Rebecq et al., 2019) model on ESIM-280 (Stoffregen et al., 2020) using the publicly available V2V training pipeline, and evaluate on HQF (Stoffregen et al., 2020), IJRR (Mueggler et al., 2017), and MVSEC (Zhu et al., 2018). Reconstruction quality is assessed using MSE, SSIM, and LPIPS. We use `T30-B300-D5` throughout. Results from methods without the same public training pipeline are included for reference only.

Table 3 shows that our learned latent consistently outperforms both interpolated and discrete voxel grids under the same V2V-E2VID training protocol across IJRR, MVSEC, and HQF. Compared with Voxel-Disc., it reduces MSE from 0.068 to 0.043, 0.147 to 0.089, and 0.041 to 0.038, respectively, while also improving SSIM and LPIPS on all three datasets. Among the reported results, our V2V-E2VID variant also achieves the second-best MSE, SSIM, and LPIPS on MVSEC, although these cross-architecture comparisons are provided for reference because the training pipelines differ.



(a) Training Dynamics (b) Pixel-level Reconstruction (c) Scene-level Reconstruction

Figure 8: **Effect of representation structure on latent learning.** Comparison between the sparse conflict-free volume $\mathbf{V}$ and the physics-grounded proxy $\mathbf{P}$.

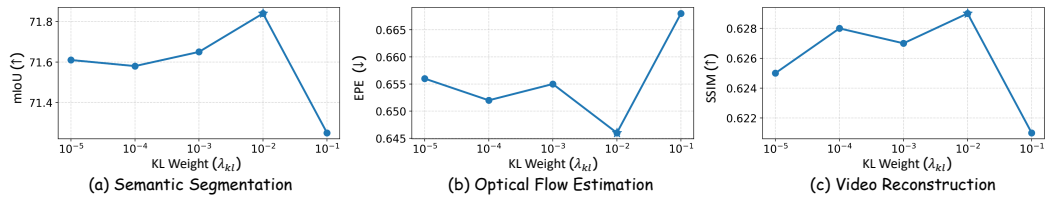
4.3 Ablation Studies

4.3.1 Event Representation

To assess the effect of forward-physics reorganization, we compare latent learning from the sparse volume $\mathbf{V}$ and the proxy $\mathbf{P}$ under the same **`T30-B300-D3`** configuration and training protocol. Consistent with the preceding analysis, Fig. 8 shows that, under this matched setting, direct learning from $\mathbf{V}$ exhibits unstable optimization and lower reconstruction fidelity, whereas $\mathbf{P}$ supports stable training and more faithful reconstruction. This comparison indicates that forward-physics reorganization improves the suitability of the event representation for latent compression.

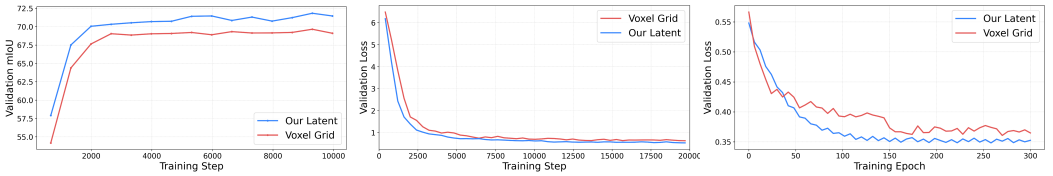
4.3.2 Downstream Tasks

We further examine how latent regularization affects downstream transfer. As shown in Fig. 9, moderate KL regularization provides the most consistent performance across all three tasks, with $\lambda_{\text{kl}} = 10^{-2}$ used as the default setting. Consistent with Fig. 6, increasing λ_{kl} beyond 10^{-2} reduces reconstruction fidelity, accompanied by a decline in downstream performance. Fig. 10 further shows stable and generally faster optimization with the proposed latent than with voxel-grid inputs across the three downstream tasks.



(a) Semantic Segmentation (b) Optical Flow Estimation (c) Video Reconstruction

Figure 9: **Effect of KL regularization on downstream transfer.** A moderate weight of $\lambda_{\text{kl}} = 10^{-2}$ yields the most consistent performance across tasks.



(a) Semantic Segmentation (b) Optical Flow Estimation (c) Video Reconstruction

Figure 10: **Downstream training dynamics.** Compared with voxel-grid inputs, the proposed latent exhibits stable and generally faster convergence across the three downstream tasks.

5 Conclusion and Discussion

We revisited a fundamental question in event representation learning: how can irregular event signals be structured and compressed while preserving the temporal information they encode? To this end, we presented a fidelity-aware framework for learning compact, statistically tractable, and transferable event representations. Our results show that reorganizing event information into more coherent forms enables faithful compression, producing latents that retain event dynamics while generalizing across domains and transferring effectively to diverse visual tasks. More broadly, this work suggests a simple principle for event representation learning: reorganize event information according to sensing physics before compression.

AI USE STATEMENT

In this work, we did not use generative AI tools for research ideation, hypothesis formulation, methodology design, implementation of the proposed method, experimental analysis, or interpretation of the results. Generative AI tools were used only for language editing, including improving grammar, clarity, and readability. All AI-assisted edits were carefully reviewed by the authors to ensure that they did not alter the technical meaning or scientific claims of the paper. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive data. We are not aware of any significant ethical concerns arising from the methodology or experiments presented in this paper.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, experimental settings, datasets, evaluation protocols, and implementation details in the main paper and appendix. Hyperparameter settings and additional experimental details are also included to facilitate reproduction of our results. We plan to release the source code upon publication.

REFERENCES

Mohammed Almatrafi, Raymond Baldwin, Kiyoharu Aizawa, and Keigo Hirakawa. Distance surface for event-based optical flow. *IEEE transactions on pattern analysis and machine intelligence*, 42 (7):1547–1556, 2020.

Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Ignacio Alzugaray and Margarita Chli. Asynchronous corner detection and tracking for event cameras in real time. *IEEE Robotics and Automation Letters*, 3(4):3177–3184, 2018.

Lakshmi Annamalai, Vignesh Ramanathan, and Chetan Singh Thakur. Event-lstm: An unsupervised and asynchronous learning-based representation for event-based data. *IEEE Robotics and Automation Letters*, 7(2):4678–4685, 2022.

R Wes Baldwin, Ruixu Liu, Mohammed Almatrafi, Vijayan Asari, and Keigo Hirakawa. Time-ordered recent event (tore) volumes for event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2519–2532, 2022.

Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based spatio-temporal feature learning for neuromorphic vision sensing. *IEEE Transactions on Image Processing*, 29:9084–9098, 2020.

Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd17: End-to-end davis driving dataset. *arXiv preprint arXiv:1711.01458*, 2017.

Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *European Conference on Computer Vision*, pp. 136–152. Springer, 2020.

Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M. Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J. Taylor, and Kostas Daniilidis. M3ed: Multi-robot, multi-sensor, multi-environment event dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4015–4022, June 2023.

Jingxi Chen, Brandon Y Feng, Haoming Cai, Tianfu Wang, Levi Burner, Dehao Yuan, Cornelia Fermuller, Christopher A Metzler, and Yiannis Aloimonos. Repurposing pre-trained video diffusion models for event-based video interpolation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12456–12466, 2025a.

Junming Chen, Jingjing Meng, Xinchao Wang, and Junsong Yuan. Dynamic graph cnn for event-camera based gesture recognition. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5. IEEE, 2020.

Kanghao Chen, Zixin Zhang, Guoqiang Liang, Lutao Jiang, Zeyu Wang, and Ying-Cong Chen. Event-guided consistent video enhancement with modality-adaptive diffusion pipeline. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Nuo Chen, Boyang Li, Yingqian Wang, Xinyi Ying, Longguang Wang, Chushu Zhang, Yulan Guo, Miao Li, and Wei An. Motion and appearance decoupling representation for event cameras. *IEEE Transactions on Image Processing*, 2025c.

Zhiwen Chen, Jinjian Wu, Junhui Hou, Leida Li, Weisheng Dong, and Guangming Shi. Ecsnet: Spatio-temporal feature learning for event camera. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2):701–712, 2022.

Zhiwen Chen, Zhiyu Zhu, Yifan Zhang, Junhui Hou, Guangming Shi, and Jinjian Wu. Segment any event streams via weighted adaptation of pivotal tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3890–3900, 2024.

Zhiwen Chen, Junhui Hou, Zhiyu Zhu, Jinjian Wu, and Guangming Shi. Scaling dense event-stream pretraining from visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Tobi Delbrück, Bernabe Linares-Barranco, Eugenio Culurciello, and Christoph Posch. Activity-driven, event-based vision sensors. In *Proceedings of 2010 IEEE international symposium on circuits and systems*, pp. 2426–2429. IEEE, 2010.

Yongjian Deng, Hao Chen, Hai Liu, and Youfu Li. A voxel graph cnn for object classification with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1172–1181, 2022.

P Kingma Diederik and Welling Max. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Burak Ercan, Onur Eker, Canberk Saglam, Aykut Erdem, and Erkut Erdem. Hypere2vid: Improving event-based video reconstruction via hypernetworks. *IEEE Transactions on Image Processing*, 33: 1826–1837, 2024.

Rui Fan, Weidong Hao, Juntao Guan, Lai Rui, Lin Gu, Tong Wu, Fanhong Zeng, and Zhangming Zhu. Eventpillars: Pillar-based efficient representations for event data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2861–2869, 2025.

Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5633–5643, 2019.

Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3586–3595, 2020.

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021a.

Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *2021 International Conference on 3D Vision (3DV)*, pp. 197–206. IEEE, 2021b.

Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense continuous-time optical flow from event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7): 4736–4746, 2024.

Ruipeng Guo, Qianwan Yang, Andrew S Chang, Guorong Hu, Joseph Greene, Christopher V Gabel, Sixian You, and Lei Tian. Eventlfm: event camera integrated fourier light field microscopy for ultrafast 3d imaging. *Light: Science & Applications*, 13(1):144, 2024.

Ondřej Holešovskỳ, Radoslav Škoviera, Václav Hlaváč, and Roman Vitek. Experimental comparison between event and global shutter cameras. *Sensors*, 21(4):1137, 2021.

Rui Hu, Song Wu, Wen Yang, and Jinjian Wu. From contrast to consistency: Rethinking event-based continuous-time optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15125–15134, June 2026.

Zexi Jia, Kaichao You, Weihua He, Yang Tian, Yongxiang Feng, Yaoyuan Wang, Xu Jia, Yihang Lou, Jingyi Zhang, Guoqi Li, et al. Event-based semantic segmentation with posterior attention. *IEEE Transactions on Image Processing*, 32:1829–1842, 2023.

Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2146–2156, 2021.

Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18032–18042, June 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lingdong Kong, Youquan Liu, Lai Xing Ng, Benoit R Cottereau, and Wei Tsang Ooi. Openess: Event-based semantic scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15686–15698, 2024.

Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016.

Ke Li, Gengyu Lyu, Hao Chen, Bochen Xie, Zhen Yang, Youfu Li, and Yongjian Deng. Know where you are from: Event-based segmentation via spatio-temporal propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4806–4814, 2025.

Yijin Li, Han Zhou, Bangbang Yang, Ye Zhang, Zhaopeng Cui, Hujun Bao, and Guofeng Zhang. Graph-based asynchronous event processing for rapid object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 934–943, 2021.

Quanmin Liang, Qiang Li, Shuai Liu, Xinzi Cao, Jinyi Lu, Feidiao Yang, Wei Zhang, Kai Huang, and Yonghong Tian. Efficient event camera data pretraining with adaptive prompt fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8656–8667, 2025.

Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128 x 128 120db 30mw asynchronous vision sensor that responds to relative intensity change. In *2006 IEEE International Solid State Circuits Conference-Digest of Technical Papers*, pp. 2060–2069. IEEE, 2006.

Songnan Lin, Ye Ma, Jing Chen, and Bihan Wen. Compressed event sensing (ces) volumes for event cameras. *International Journal of Computer Vision*, 133(1):435–455, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Bingde Liu, Chang Xu, Wen Yang, Huai Yu, and Lei Yu. Motion robust high-speed light-weighted object detection with event camera. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–13, 2023a.

Daikun Liu, Lei Cheng, Teng Wang, and Changyin Sun. Edcflow: Exploring temporally dense difference maps for event-based optical flow estimation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1984–1993. IEEE, 2025a.

Haotian Liu, Guang Chen, Sanqing Qu, Yanping Zhang, Zhijun Li, Alois Knoll, and Changjun Jiang. Tma: Temporal motion aggregation for event-based optical flow. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9651–9660. IEEE, 2023b.

Haoyue Liu, Shihan Peng, Lin Zhu, Yi Chang, Hanyu Zhou, and Luxin Yan. Seeing motion at nighttime with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25648–25658, 2024.

Haoyue Liu, Jinghan Xu, Shihan Peng, Yi Chang, Hanyu Zhou, Yuxing Duan, Lin Zhu, Yonghong Tian, and Luxin Yan. Ner-net+: Seeing motion at nighttime with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4768–4786, 2025b. doi: 10.1109/TPAMI.2025.3545936.

Hanyue Lou, Jinxiu Liang, Minggui Teng, Yi Wang, and Boxin Shi. V2v: Scaling event-based vision through efficient video-to-voxel simulation. *Advances in Neural Information Processing Systems*, 38:122485–122517, 2026.

Xinglong Luo, Kunming Luo, Ao Luo, Zhengning Wang, Ping Tan, and Shuaicheng Liu. Learning optical flow from event camera with rendered dataset. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9813–9823. IEEE, 2023.

Xinglong Luo, Ao Luo, Zhengning Wang, Chunyu Lin, Bing Zeng, and Shuaicheng Liu. Efficient meshflow and optical flow estimation from event cameras. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19198–19207. IEEE, 2024.

Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide 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*, pp. 5419–5427, 2018.

Balázs Mészáros, James C Knight, and Thomas Nowotny. Efficient event-based delay learning in spiking neural networks. *Nature Communications*, 16(1):10422, 2025.

Mohammad Mostafavi, Lin Wang, and Kuk-Jin Yoon. Learning to reconstruct hdr images from events, with applications to depth and flow prediction. *International Journal of Computer Vision*, 129(4):900–920, 2021.

Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International journal of robotics research*, 36(2):142–149, 2017.

Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Yeongwoo Nam, Mohammad Mostafavi, Kuk-Jin Yoon, and Jonghyun Choi. Stereo depth from events cameras: Concentrate and focus on the future. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6114–6123, 2022.

Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. *IEEE Journal of Solid-State Circuits*, 46(1):259–275, 2010.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

Bharath Ramesh, Hong Yang, Garrick Orchard, Ngoc Anh Le Thi, Shihao Zhang, and Cheng Xiang. Dart: distribution aware retinal transform for event-based cameras. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2767–2780, 2019.

Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization. In *BMVC*, volume 2, pp. 7, 2017.

Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on robot learning*, pp. 969–982. PMLR, 2018.

Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3852–3861. IEEE, 2019.

Hongwei Ren, Yue Zhou, Haotian Fu, Yulong Huang, Renjing Xu, and Bojun Cheng. Ttpoint: A tensorized point cloud network for lightweight action recognition with event cameras. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8026–8034, 2023.

Hongwei Ren, Yue Zhou, Jiadong Zhu, Xiaopeng Lin, Haotian Fu, Yulong Huang, Yuetong Fang, Fei Ma, Hao Yu, and Bojun Cheng. Rethinking efficient and effective point-based networks for event camera classification and regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1731–1740, 2018.

Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *European Conference on Computer Vision*, pp. 534–549. Springer, 2020.

Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pp. 341–357. Springer, 2022.

Minggui Teng, Chu Zhou, Hanyue Lou, and Boxin Shi. Nest: Neural event stack for event-based image enhancement. In *European Conference on Computer Vision*, pp. 660–676. Springer, 2022.

Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10081–10090, 2019a.

Lin Wang, Yujeong Chae, and Kuk-Jin Yoon. Dual transfer learning for event-based end-task prediction via pluggable event to image translation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2135–2145, 2021a.

Lin Wang, Yujeong Chae, Sung-Hoon Yoon, Tae-Kyun Kim, and Kuk-Jin Yoon. Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 608–619, 2021b.

Qinyi Wang, Yexin Zhang, Junsong Yuan, and Yilong Lu. Space-time event clouds for gesture recognition: From rgb cameras to event cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1826–1835. IEEE, 2019b.

Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9700–9709, 2021c.

Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2543–2552. IEEE, 2021.

Yilun Wu, Federico Paredes-Vallés, and Guido CHE De Croon. Lightweight event-based optical flow estimation via iterative deblurring. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14708–14715. IEEE, 2024.

Xinan Xie, Qing Zhang, and Wei-Shi Zheng. Diffusion-based event generation for high-quality image deblurring. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2194–2203, 2025.

Gang Xu, Zhiyu Zhu, and Junhui Hou. Unie2f: A unified diffusion framework for event-to-frame reconstruction with video foundation models. *arXiv preprint arXiv:2602.19202*, 2026.

Jiancheng Yang, Qiang Zhang, Bingbing Ni, Linguo Li, Jinxian Liu, Mengdie Zhou, and Qi Tian. Modeling point clouds with self-attention and gumbel subset sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3323–3332, 2019.

Yan Yang, Liyuan Pan, and Liu Liu. Event camera data dense pre-training. In *European Conference on Computer Vision*, pp. 292–310. Springer, 2024.

Yixin Yang, Jiawei Zhang, Yang Zhang, Yunxuan Wei, Dongqing Zou, Jimmy S Ren, and Boxin Shi. Event-guided hdr reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11787–11796, 2025.

Man Yao, Huanhuan Gao, Guangshe Zhao, Dingheng Wang, Yihan Lin, Zhaoxu Yang, and Guoqi Li. Temporal-wise attention spiking neural networks for event streams classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10221–10230, 2021.

Bohan Yu, Jieji Ren, Jin Han, Feishi Wang, Jinxiu Liang, and Boxin Shi. Eventps: Real-time photometric stereo using an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9602–9611, 2024.

Yucheng Zhao, Gengyu Lyu, Ke Li, Zihao Wang, Hao Chen, Zhen Yang, and Yongjian Deng. Eseg: Event-based segmentation boosted by explicit edge-semantic guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10510–10518, 2025.

Qianang Zhou, Zhiyu Zhu, Junhui Hou, Yongjian Deng, Youfu Li, and Junlin Xiong. Resflow: Fine-tuning residual optical flow for event-based high temporal resolution motion estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Alex Zihao Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018.

Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 989–997, 2019.

Jinjing Zhu, Tianbo Pan, Zidong Cao, Yexin Liu, James T Kwok, and Hui Xiong. Depth any event stream: Enhancing event-based monocular depth estimation via dense-to-sparse distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5146–5155, 2025.

Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, and Yonghong Tian. Event-based video reconstruction via potential-assisted spiking neural network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3594–3604, 2022a.

Zhiyu Zhu, Junhui Hou, and Xianqiang Lyu. Learning graph-embedded key-event back-tracing for object tracking in event clouds. *Advances in Neural Information Processing Systems*, 35: 7462–7476, 2022b.

Nikola Zubić, Daniel Gehrig, Mathias Gehrig, and Davide Scaramuzza. From chaos comes order: Ordering event representations for object recognition and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12846–12856, 2023.

Nikola Zubic, Mathias Gehrig, and Davide Scaramuzza. State space models for event cameras. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5819–5828, 2024.