ROBUSTNESS FROM INTERACTION: EVENT-STRUCTURE-AWARE ATTENTION FOR EVENT-BASED SPIKING TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Robustness in event-based vision is often addressed before the network via event denoising, corruption-aware training schemes, or robust feature representations. We pose a complementary question: after noisy events have already been discretized and tokenized, can we improve robustness by controlling how event tokens interact with one another? Unlike generic sequence tokens, event tokens still encode sensing-imposed spatial and polarity structure inherited from the event generation process. We introduce Spiking Event-Consistent Attention (SECA), which maps this token-aligned structure into a bounded multiplicative modulation of learned query–key correlations, adding only two learnable scalars per block while keeping spiking dynamics intact. Without any corruption training (trained only on clean data), SECA maintains clean accuracy across three Spiking Transformer families and cuts the performance drop under the strongest polarity flip on DVS128-Gesture from 28.82 to 9.38 points, with improvements in most background activity and hot-pixel noise settings. Controlled violations of the assumed priors indicate that retaining event structure provides benefits beyond a generic bounded modulation, especially under polarity corruption. A test-time gate ablation further shows that much of the robustness remains even when the gate is turned off at inference, suggesting the structural constraint primarily shapes the learned solution rather than serving only as test-time suppression. Overall, these findings position interaction-level robustness as a complementary design principle for event-based Spiking Transformers.

1 INTRODUCTION

Event cameras produce sparse, asynchronous measurements of local brightness changes with signed polarity, providing a natural substrate for event-driven perception Gallego et al. (2022). Recent Spiking Transformers further enable global interactions over event-derived representations Zhang et al. (2022); Zhou et al. (2023); Yao et al. (2023; 2024). Robustness to sensor and stream corruption is still typically pursued *upstream* of these interactions, through event filtering or denoising and robust representation learning Guo & Delbruck (2023); Wan et al. (2022); Jiang et al. (2024); Duan (2024). Such approaches largely intervene before or beneath the token-interaction mechanism. We therefore ask a complementary question: *once corrupted events have been tokenized, can robustness be induced by regulating how event tokens interact?* This motivates an interaction-level view of event robustness within the learned attention process itself.

Event tokens are not generic sequence elements: they aggregate threshold-triggered brightness changes over localized space and time, and thus retain sensing-derived spatial and polarity structure after discretization and tokenization Lichtsteiner et al. (2008); Gallego et al. (2022); Schaefer et al. (2022); Peng et al. (2023); Wu et al. (2024). Local motion induces spatially structured responses, while polarity encodes the direction of brightness change Gallego et al. (2022); Cao et al. (2026). Although feature-driven attention may capture such cues implicitly, its query–key correlation does not explicitly model the spatial–polarity relation between interacting tokens at the point where pairwise interactions are formed Zhou et al. (2023); Peng et al. (2023). This motivates a simple principle: *event-token interactions should remain softly compatible with sensing-derived structure, and such compatibility may itself provide robustness.*

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.

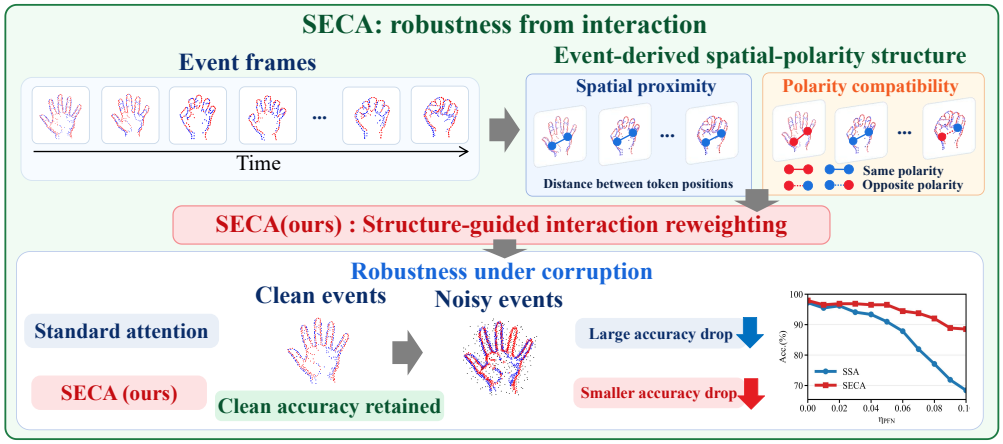


Figure 1: Motivation and overview of SECA. SECA uses spatial–polarity structure to reweight interactions, preserving clean accuracy while improving robustness over standard attention.

As conceptually illustrated in Fig. 1, we instantiate this principle with a deliberately minimal mechanism. The resulting Spiking Event-Consistent Attention (SECA) constructs an event-derived, token-aligned spatial–polarity prior, $\Phi_{ij} = \rho_i \rho_j \exp(-d_{ij}^2/\sigma_s^2)$, and uses it to modulate the query–key interaction of self-attention Vaswani et al. (2017), specifically the Spiking Self-Attention (SSA) correlation Zhou et al. (2023), $S_{ij}^{\mathrm{SECA}} = S_{ij}^{\mathrm{SSA}}(1 + \lambda \Phi_{ij})$. With $\lambda = \frac{1}{2}\mathrm{Sigmoid}(\lambda_{\mathrm{raw}})$, the modulation is bounded by $1 + \lambda\Phi_{ij} \in (0.5, 1.5)$, so the prior only reweights interactions already supported by learned features rather than replacing learned feature correlations. SECA adds only two learnable scalars per attention block, leaves the surrounding spiking dynamics unchanged, and requires neither a denoising module nor corruption-aware training.

Despite its minimal intervention, SECA largely preserves clean-data recognition across three Spiking Transformer families and multiple event datasets, while substantially improving robustness when trained only on clean data. On DVS128-Gesture Amir et al. (2017) with Spikformer, the clean-relative accuracy drop under the strongest Polarity Flip Noise (PFN) decreases from 28.82 to 9.38 percentage points, with additional gains under Background Activity Noise (BAN) and Hot-Pixel Noise (HPN). Controlled prior violations show that these gains cannot be explained solely by generic bounded modulation, particularly under PFN and HPN. Moreover, disabling the SECA gate only at inference preserves most of the PFN robustness, suggesting that the structural bias may also shape the solution learned during training rather than acting only as test-time suppression. Taken together, these findings lead to three main contributions:

(i) An interaction-level view of event robustness. We formulate token interaction as a complementary level of robustness intervention, alongside input- and representation-level approaches, and identify event-derived structural compatibility as a principled cue for regulating interactions once corrupted events have been tokenized (Sec. 2).

(ii) A minimal instantiation through event-consistent attention. We propose SECA, which injects a bounded, token-aligned spatial–polarity prior into spiking attention using only two learnable scalars per block, while retaining learned feature correlation as the primary interaction signal and leaving the host spiking dynamics unchanged (Sec. 3).

(iii) A falsifiable analysis of the robustness mechanism. Beyond standard corruption benchmarks, we use controlled prior violations to test whether event structure and token alignment matter, analyze modulation under corruption, and perform a test-time gate ablation to examine whether the learned robustness can be explained solely by instantaneous inference-time suppression (Secs. 4.4–4.6).

2 Related Work

Event-based vision and event representation. Event cameras asynchronously encode per-pixel brightness changes with timestamps and polarity, yielding sparse streams with high temporal reso-

lution and dynamic range (Gallego et al., 2022). Because these streams are irregular, event-based learning commonly relies on structured representations such as event frames, time surfaces, voxel grids, or learned encodings (Gehrig et al., 2019; Teng et al., 2022; Zubić et al., 2023; Zubic et al., 2024). Recent work increasingly develops task-adaptive representations that preserve or exploit event-specific temporal, spatial, and polarity structure, including multimodal event embeddings, frequency-aware representations, spatiotemporal multi-view encodings, and learnable event aggregation (Peng et al., 2023; Zhou et al., 2024b; Cao et al., 2025; 2026; Fan et al., 2026; Xue et al., 2026). Event streams are also affected by background activity, illumination variation, and sensor artifacts, motivating denoising, robust representation learning, and domain adaptation before or during feature construction (Guo & Delbruck, 2023; Wan et al., 2022; Jiang et al., 2024; Duan, 2024; Burkhardt et al., 2025; Kong et al., 2025). SECA instead operates after tokenization, using sensing-derived spatial–polarity structure to regulate token interactions rather than modifying the input stream or its representation.

Spiking Transformers and spiking attention. Spiking Transformers combine spike-based temporal dynamics with Transformer-style global interaction. Spikformer introduced SSA with spike-form queries, keys, and values (Zhou et al., 2023), while Spike-driven Transformer developed Spike-Driven Self-Attention (SDSA) for efficient spike-driven computation (Yao et al., 2023). Subsequent work has further improved scalability, efficiency, and spatiotemporal interaction through hierarchical Q–K attention, expert routing, large-scale training, addition-only computation, token pruning, and spatial–temporal attention (Yao et al., 2024; Zhou et al., 2024a;c; Yao et al., 2025; Guo et al., 2025; Lee et al., 2025; Wei et al., 2026). More closely related to our focus, Neural Dynamics Self-Attention introduces locality bias into spiking attention (Zhang et al., 2026), while Bipolar Self-Attention models homopolar and heteropolar interactions from signed membrane-potential representations (Wang et al., 2025). SECA differs by deriving an input-dependent pairwise prior jointly from sensor ON/OFF polarity and token geometry, and using this sensing-derived spatial–polarity compatibility as a bounded modulation of learned token interactions for robustness under event corruption, while preserving the backbone's original spiking computation and temporal dynamics.

Robustness interventions. We organize existing event-robustness strategies by their *primary intervention point*. *Input-level* approaches clean or refine event streams through spatiotemporal filtering, learned denoising, joint motion–noise estimation, or task-aware event refinement (Guo & Delbruck, 2023; Jiang et al., 2024; Duan, 2024; Shiba et al., 2025; Chen et al., 2026; Yao et al., 2026). *Representation-level* approaches instead improve robustness through robust feature learning, synthetic domain translation, feature alignment, or domain adaptation (Wan et al., 2022; Jeong et al., 2024; Kong et al., 2025; Hamann et al., 2025; Li et al., 2025; Xu et al., 2026). Although these categories are not mutually exclusive, their primary interventions remain upstream of the token-interaction mechanism. SECA instead targets a complementary *interaction level*: it directly modulates how already-tokenized events interact using sensing-derived spatial–polarity compatibility, without additional input preprocessing or corruption-aware training. Rather than replacing input denoising or robust representation learning, SECA modifies the token-interaction operator while preserving the surrounding spiking dynamics and the clean-only training protocol.

3 METHOD

The interaction-level view does not prescribe a unique mechanism. We therefore adopt three design requirements: (i) learned feature correlation remains the primary interaction signal; (ii) event-derived structure modulates, rather than replaces, existing interactions; and (iii) the intervention introduces negligible additional capacity. These requirements motivate SECA, a lightweight bounded modulation of spiking attention, as illustrated in Fig. 2.

3.1 PROBLEM SETUP AND TOKEN-ALIGNED EVENT REPRESENTATION

An event camera asynchronously records local brightness changes as a stream of events (Gallego et al., 2022),

$$\mathcal{E} = \{e_m = (x_m, y_m, \tau_m, p_m)\}_{m=1}^{M}, \qquad p_m \in \{-1, +1\}, \tag{1}$$

where (x_m, y_m), τ_m, and p_m denote the spatial location, timestamp, and polarity of the m-th event, respectively. Following the discrete-time processing commonly adopted in Spiking Trans-

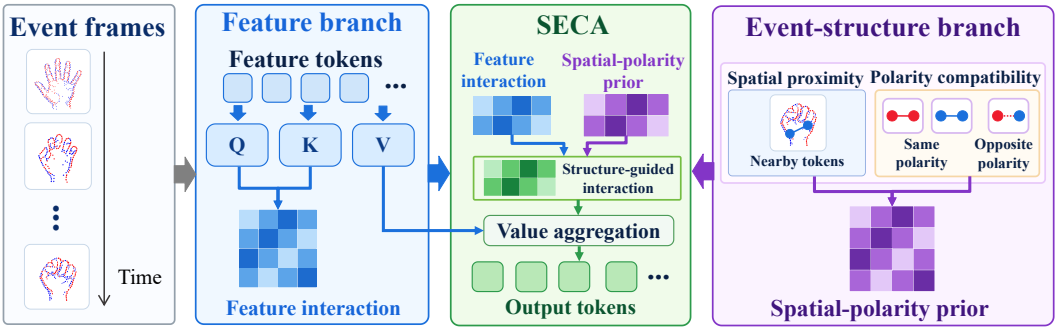


Figure 2: Overview of SECA. The feature branch produces query–key interactions, while the event-structure branch derives a token-aligned spatial–polarity prior from spatial proximity and polarity compatibility. SECA uses this prior to reweight interactions before value aggregation.

formers (Zhou et al., 2023; Yao et al., 2023), the event stream is partitioned into T temporal steps and converted into spatial tokens. We denote the resulting feature tensor and token-aligned polarity descriptor as

$$\mathbf{X} \in \mathbb{R}^{T\times B\times C\times N}, \qquad \boldsymbol{\rho} \in [-1,1]^{T\times B\times N}, \tag{2}$$

where B, C, and N denote the batch size, feature dimension, and number of spatial tokens, respectively. Each token represents an aggregated spatial region within one temporal step rather than an individual raw event. We therefore distinguish the binary raw-event polarity p_m from the continuous token-aligned polarity descriptor $\rho_{t,b,n}$, which aggregates polarity information for token n at step t. A formal description of the construction and alignment of $\boldsymbol{\rho}$ is provided in Appendix A.

The $N = H_p^2$ tokens form a regular $H_p \times H_p$ grid. Each token n is assigned a normalized coordinate $\mathbf{c}_n = (u_n, v_n) \in [0,1]^2$, and the squared distance between tokens i and j is $d_{ij}^2 = \|\mathbf{c}_i - \mathbf{c}_j\|_2^2$. This normalized geometry makes the spatial relation independent of the absolute token-grid resolution.

For H attention heads with head dimension $d_h = C/H$, the spike-form projections produce

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{T\times B\times H\times N\times d_h}.$$

The pairwise SSA correlation (Zhou et al., 2023) at temporal step t, batch element b, and head h is

$$\mathbf{S}^{\mathrm{SSA}}_{t,b,h} = \mathbf{Q}_{t,b,h}\mathbf{K}^{\top}_{t,b,h} \in \mathbb{R}^{N\times N}. \tag{3}$$

Although learned features may implicitly encode event structure, the interaction score in equation 3 contains no explicit pairwise term for token geometry or event polarity. SECA introduces this sensing-derived structural cue while retaining feature correlation as the primary interaction signal.

3.2 Event-Derived Spatial–Polarity Prior

The event-generation mechanism motivates spatial and polarity information as complementary structural cues. Under the standard event-camera model (Lichtsteiner et al., 2008; Gallego et al., 2022; Liu et al., 2024), an event is triggered when the local log-intensity change reaches a contrast threshold, $\Delta L_m \approx p_m \theta_{\mathrm{th}}$, where $\theta_{\mathrm{th}} > 0$. Thus, event polarity reflects the sign of the threshold-crossing brightness change and is tied directly to the sensing process. We use this relation only to motivate a *soft structural prior*: a particular polarity configuration is not assumed to uniquely determine whether two events arise from the same physical process. A local-motion interpretation and the scope of this physical motivation are provided in Appendix B.

Using the normalized token coordinates defined in Sec. 3.1, we model spatial locality as

$$\kappa_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma_s^2}\right), \qquad \sigma_s^2 = \exp(\alpha_s), \tag{4}$$

where α_s is a learnable scalar. Smaller σ_s^2 emphasizes local interactions, whereas larger values yield a broader spatial range.

At temporal step t, the polarity relation between tokens i and j is represented by $\rho_{t,b,i}\rho_{t,b,j}$. Combining this relation with spatial locality gives the proposed spatial–polarity prior,

$$\Phi_{t,b,i,j} = \rho_{t,b,i}\rho_{t,b,j}\kappa_{ij} = \rho_{t,b,i}\rho_{t,b,j}\exp\left(-\frac{d_{ij}^2}{\sigma_s^2}\right). \tag{5}$$

Equivalently, $\mathbf{\Phi}_{t,b} = (\boldsymbol{\rho}_{t,b}\boldsymbol{\rho}_{t,b}^\top) \odot \boldsymbol{\kappa}$, where $\boldsymbol{\rho}_{t,b} \in [-1,1]^N$ and $\boldsymbol{\kappa} \in (0,1]^{N\times N}$. Therefore, $\mathbf{\Phi}_{t,b} \in [-1,1]^{N\times N}$.

The prior has three useful properties. It is *token-aligned*, since $\Phi_{t,b,i,j}$ corresponds directly to the same token pair (i,j) as the attention interaction; *time-conditioned*, because $\boldsymbol{\rho}_{t,b}$ and hence $\mathbf{\Phi}_{t,b}$ vary with the event configuration at each temporal step; and *feature-independent*, because it is constructed from event-derived polarity and token geometry rather than learned query or key features. We therefore treat $\mathbf{S}_{t,b,h}^{\mathrm{SSA}}$ and $\mathbf{\Phi}_{t,b}$ as complementary feature evidence and sensing-derived structural evidence, respectively, during pairwise interaction formation at each temporal step.

3.3 Spiking Event-Consistent Attention

Given the spatial–polarity prior $\mathbf{\Phi}_{t,b} \in [-1,1]^{N\times N}$, SECA injects sensing-derived structure directly into the learned pairwise interaction through a multiplicative gate,

$$\mathbf{G}_{t,b} = \mathbf{1}_{N\times N} + \lambda\mathbf{\Phi}_{t,b}, \qquad \lambda = \frac{1}{2}\,\mathrm{Sigmoid}(\lambda_{\mathrm{raw}}), \tag{6}$$

where λ_{raw} is learnable and $\mathbf{G}_{t,b}$ is shared across attention heads. The SECA correlation is then

$$\mathbf{S}_{t,b,h}^{\mathrm{SECA}} = \mathbf{S}_{t,b,h}^{\mathrm{SSA}} \odot \mathbf{G}_{t,b} = \mathbf{S}_{t,b,h}^{\mathrm{SSA}} \odot (\mathbf{1}_{N\times N} + \lambda\mathbf{\Phi}_{t,b}). \tag{7}$$

Thus, each head retains its independently learned query–key correlation, while the same event-derived compatibility prior modulates the corresponding token pairs. Element-wise,

$$S_{t,b,h,i,j}^{\mathrm{SECA}} = S_{t,b,h,i,j}^{\mathrm{SSA}}\,(1 + \lambda\Phi_{t,b,i,j}). \tag{8}$$

Positive $\Phi_{t,b,i,j}$ mildly amplifies an existing interaction, negative values attenuate it, and values near zero leave it unchanged. Importantly, the prior cannot create an interaction independently of the learned representation: if $S_{t,b,h,i,j}^{\mathrm{SSA}} = 0$, then $S_{t,b,h,i,j}^{\mathrm{SECA}} = 0$ regardless of $\Phi_{t,b,i,j}$.

The modulation is deliberately bounded. Since $\lambda \in (0, 0.5)$ and $\Phi_{t,b,i,j} \in [-1,1]$, the gate satisfies

$$0.5 < G_{t,b,i,j} = 1 + \lambda\Phi_{t,b,i,j} < 1.5. \tag{9}$$

Because $G_{t,b,i,j} > 0$, SECA preserves the sign of every nonzero SSA correlation,

$$\mathrm{sign}\left(S_{t,b,h,i,j}^{\mathrm{SECA}}\right) = \mathrm{sign}\left(S_{t,b,h,i,j}^{\mathrm{SSA}}\right). \tag{10}$$

Moreover, for $S_{t,b,h,i,j}^{\mathrm{SSA}} \neq 0$,

$$0.5\left|S_{t,b,h,i,j}^{\mathrm{SSA}}\right| < \left|S_{t,b,h,i,j}^{\mathrm{SECA}}\right| < 1.5\left|S_{t,b,h,i,j}^{\mathrm{SSA}}\right|. \tag{11}$$

Hence, the event-derived prior provides only a bounded correction to learned-feature interactions; it cannot reverse their sign or replace them with an unconstrained structural score.

The modulated correlation is subsequently used for value aggregation,

$$\mathbf{Z}_{t,b,h} = \frac{1}{\sqrt{d_h}}\mathbf{S}_{t,b,h}^{\mathrm{SECA}}\mathbf{V}_{t,b,h}, \tag{12}$$

without softmax normalization. The outputs of all heads are concatenated and passed through the original post-attention spiking and projection operations of the host block. SECA therefore modifies the token-interaction rule while preserving the surrounding temporal neuronal dynamics and feature transformations, introducing no additional state variables or changes to the spiking pipeline.

SECA introduces only two prior-specific learnable scalars per attention block: α_s, which controls the spatial scale σ_s^2 in equation 4, and λ_{raw}, which controls the modulation strength in equation 6. For a fixed token resolution, the pairwise distance matrix $\{d_{ij}^2\}_{i,j=1}^N$ can be cached. Constructing $\mathbf{\Phi}$ requires $\mathcal{O}(TBN^2)$ operations, while applying the shared gate across H heads requires $\mathcal{O}(TBHN^2)$

Table 1: Robustness across three backbones on DVS128-Gesture. Values are clean-relative ΔAcc (p.p.) for Base/SECA, reported as mean $\pm$ std. over three paired corruption realizations.

Backbone	Noise	Mean ΔAcc Base/SECA $\downarrow$	Max ΔAcc Base/SECA $\downarrow$
Spikformer-2-256	BAN	3.56$\pm$0.18 / **3.04$\pm$0.23**	7.75$\pm$0.53 / **4.51$\pm$0.60**
	PFN	11.25$\pm$0.34 / **3.70$\pm$0.20**	28.36$\pm$0.53 / **9.26$\pm$0.87**
	HPN	17.53$\pm$0.18 / **10.29$\pm$0.40**	38.89$\pm$0.60 / **36.23$\pm$2.31**
SDT-8-384	BAN	1.55$\pm$0.23 / **1.48$\pm$0.17**	2.66$\pm$0.87 / **1.74$\pm$0.60**
	PFN	7.35$\pm$0.11 / **5.13$\pm$0.44**	15.97$\pm$0.00 / **9.38$\pm$0.92**
	HPN	**3.59$\pm$0.27** / 3.60$\pm$0.25	6.48$\pm$0.53 / **5.44$\pm$0.53**
SDT-V3-2-360	BAN	13.46$\pm$0.20 / **8.19$\pm$0.09**	34.72$\pm$0.60 / **21.76$\pm$0.53**
	PFN	**0.03$\pm$0.16** / 1.81$\pm$0.21	**1.16$\pm$0.20** / 2.55$\pm$0.20
	HPN	1.40$\pm$0.24 / **1.27$\pm$0.12**	3.59$\pm$0.72 / **2.43$\pm$1.04**

additional element-wise operations. For pairwise SSA, these operations preserve the quadratic dependence on N and introduce no additional learned projections or attention heads. For backbones with different native attention formulations, we replace the corresponding attention operator with SECA while leaving the surrounding network blocks unchanged; architecture-specific complexity is reported separately in the experiments.

Overall, SECA retains feature-driven spiking attention as the primary interaction mechanism and uses event-derived spatial–polarity structure only as a lightweight, bounded correction. Additional proofs, initialization, and implementation details are provided in Appendix C.

4 Experiments

Experimental setup. We evaluate SECA on Spikformer (Zhou et al., 2023), Spike-driven Transformer (SDT) (Yao et al., 2023), and Spike-driven Transformer-V3 (SDT-V3) (Yao et al., 2025) across DVS128-Gesture (Amir et al., 2017), CIFAR10-DVS (Li et al., 2017), N-Cars (Sironi et al., 2018), and HARDVS (Wang et al., 2024). Baseline and SECA models share identical preprocessing, training, and evaluation settings, and BAN, PFN, and HPN are applied only at test time. Full details are provided in Appendix D.

4.1 Robustness and Clean-Data Compatibility

Robustness to event corruptions. We evaluate clean-trained models under test-time BAN, PFN, and HPN, perturbing background activity, polarity, and persistent spatial responses, respectively; BAN and HPN reflect common event-camera noise (Kim et al., 2021; Guo & Delbruck, 2023; Duan, 2024), while PFN provides a controlled polarity perturbation. Robustness is measured by $\Delta\mathrm{Acc}(s) = \mathrm{Acc}_{\mathrm{clean}} - \mathrm{Acc}(s)$, where lower is better. All models use the best checkpoint from training seed 2026. Fig. 3 and Table 2 report Spikformer-2-256 results for corruption seed 2026, matching the abstract setting, while Table 1 summarizes three paired realizations. For this Spikformer realization, SECA reduces mean PFN degradation from 11.50 to 3.82 p.p. and Max-Severity degradation from 28.82 to 9.38 p.p., corresponding to 68.403% vs. 88.542% accuracy at $\eta_{\mathrm{PFN}} = 0.10$. Across three backbones, SECA lowers mean degradation in seven of nine backbone–corruption settings and Max-Severity degradation in eight of nine; the remaining cases show that robustness gains depend on both backbone and corruption type rather than being uniform. Additional severity curves and cross-dataset robustness results are provided in Appendix E.1; real-world evaluation on HARDVS is reported in Appendix E.6.

Clean-data compatibility. Table 3 shows that across 15 architecture–dataset configurations, SECA changes Top-1 accuracy by -0.20 to $+1.40$ p.p., improving 11 settings, matching one, and slightly decreasing three. Parameter counts remain unchanged at the reported precision, consistent with only two additional learnable scalars per block. These results indicate that SECA can be integrated across diverse Spiking Transformer architectures without materially compromising clean-data recognition or requiring additional model capacity. Additional analysis of temporal-resolution compatibility is provided in Appendix E.5.

Table 2: Representative corruption accuracies of Spikformer-2-256 on DVS128-Gesture.

	BAN			PFN			HPN		
Method	0.20	1.00	2.00	0.01	0.05	0.10	0.02	0.10	0.20
Base	96.875	94.097	88.889	95.486	90.972	68.403	95.833	84.722	57.986
SECA (Ours)	96.875	**94.444**	**93.750**	**96.528**	**96.528**	**88.542**	**96.875**	**95.139**	**62.847**

Table 3: Clean-data Top-1 accuracy (%) across backbones and datasets. Parentheses show SECA changes; bold indicates the better result.

Backbone	**Attention**	**Params (M)**	**CIFAR10-DVS**	**DVS128-Gesture**	**N-Cars**
Spikformer-2-256	SSA	2.58	**80.30**	97.22	94.13
	SECA (Ours)	2.58	80.20 (−0.10)	**97.92** (+0.70)	**94.16** (+0.03)
SDT-8-384	SDSA	16.43	75.00	97.57	94.69
	SECA (Ours)	16.43	**76.40** (+1.40)	**98.26** (+0.69)	**94.71** (+0.02)
SDT-8-512	SDSA	29.18	75.00	**98.96**	94.55
	SECA (Ours)	29.18	**75.60** (+0.60)	**98.96** (0.00)	**94.67** (+0.12)
SDT-V3-2-360	SDSA	5.60	**82.90**	97.92	92.67
	SECA (Ours)	5.60	82.70 (−0.20)	**98.26** (+0.34)	**92.75** (+0.08)
SDT-V3-8-360	SDSA	18.64	82.40	97.22	**92.87**
	SECA (Ours)	18.64	**83.40** (+1.00)	**98.61** (+1.39)	92.83 (−0.04)

4.2 Ablation of the Spatial–Polarity Prior

We isolate the spatial and polarity components of SECA on DVS128-Gesture using Spikformer. As shown in Table 4, both components independently improve robustness while preserving clean-data accuracy, demonstrating that each source of event-derived structure provides useful interaction constraints. The polarity-only variant shows a stronger gain under PFN, where corruption directly alters event signs and therefore challenges polarity consistency. In contrast, the spatial component provides complementary geometric information that is most helpful under BAN, where event locations and local activity patterns are disturbed. Combining both cues yields the best performance under all three corruptions while tying the best clean accuracy, supporting that spatial proximity and polarity compatibility capture different but complementary aspects of event-token structure.

Table 4: Spatial–polarity ablation on DVS128-Gesture at $\rho_{\mathrm{BAN}} = 2.0$, $\eta_{\mathrm{PFN}} = 0.10$, and $d_{\mathrm{HPN}} = 0.10$.

Variant	**Spatial**	**Polarity**	**Clean**	**BAN**	**PFN**	**HPN**
SSA	×	×	97.222	88.889	68.403	84.722
Spatial only	✓	×	**97.917**	93.056	76.389	92.361
Polarity only	×	✓	97.222	91.667	78.819	94.097
SECA	✓	✓	**97.917**	**93.750**	**88.542**	**95.139**

4.3 Attention and Representation Analysis

Figure 4 compares DVS128-Gesture Sample 24 using the same query token under clean input and PFN ($\eta_{\mathrm{PFN}} = 0.10$). SSA and SECA show similar dominant interactions on clean input, suggesting that SECA preserves the original attention pattern while adding structured modulation. Under PFN, SECA retains a pattern closer to its clean counterpart while SSA changes substantially, consistent with the smaller PFN degradation of SECA. We further visualize representations under strong BAN ($\rho_{\mathrm{BAN}} = 2.0$) using t-SNE (van der Maaten & Hinton, 2008) and UMAP (McInnes et al., 2018). In Fig. 5, red boxes (□) mark representative cross-class overlap; SSA shows greater dispersion and local class mixing, while SECA retains more coherent organization. These observations provide qualitative support for event-derived interaction guidance under corruption. Since SSA and SECA are independently trained and the embeddings are fitted independently, these visualizations are qualitative rather than causal or quantitative evidence. Additional clean, PFN, and HPN visualizations are provided in Appendix E.2.

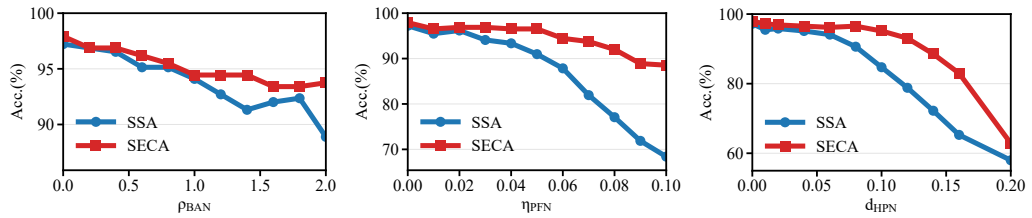
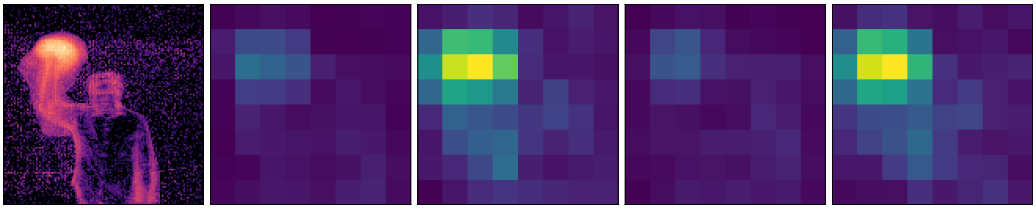


Figure 3: DVS128-Gesture robustness under BAN, PFN, and HPN (left to right).



(a) Input event (b) SSA (Clean) (c) SECA (Clean) (d) SSA ($\eta_{\mathrm{PFN}} = 0.10$) (e) SECA ($\eta_{\mathrm{PFN}} = 0.10$)

Figure 4: Attention visualization for DVS128-Gesture Sample 24 using the same query token.

4.4 Controlled Prior Violations

We test whether SECA's robustness depends on the proposed event-derived structure rather than generic bounded modulation. Shuffled-Φ permutes pairwise entries, Misaligned-Φ breaks polarity–geometry alignment, and Random-Φ removes event-specific pairwise structure while approximately preserving the prior-value distribution. As shown in Table 5, all modulated variants achieve the same clean accuracy, yet PFN degradation rises from 9.375 p.p. with the true prior to 18.056 and 24.653 p.p. for Shuffled-Φ and Misaligned-Φ, respectively, supporting the importance of token-pair correspondence and alignment. Random-Φ performs slightly better under BAN, but the true prior remains most effective under PFN and HPN, indicating that generic bounded modulation alone does not explain the robustness gain.

Table 5: Prior violations on DVS128-Gesture at $\rho_{\mathrm{BAN}} = 2.0$, $\eta_{\mathrm{PFN}} = 0.10$, and $d_{\mathrm{HPN}} = 0.10$. Results use corruption realization seed 2026.

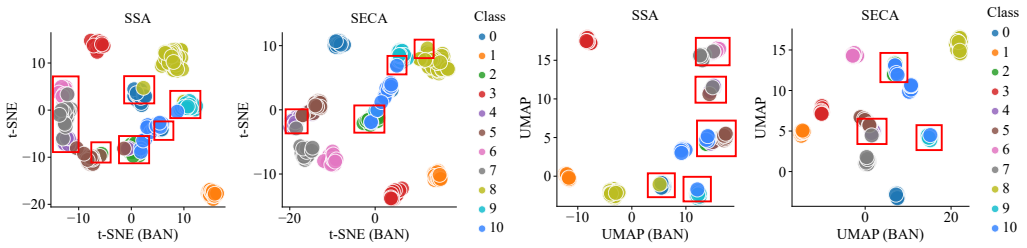
Variant	Clean Acc. (%) ↑	ΔAcc (p.p.) ↓		
		BAN	PFN	HPN
SSA	97.222	8.333	28.819	12.500
SECA	**97.917**	4.167	**9.375**	**2.778**
Shuffled-Φ	**97.917**	10.069	18.056	5.556
Misaligned-Φ	**97.917**	6.597	24.653	3.472
Random-Φ	**97.917**	**3.125**	15.278	3.472

4.5 Magnitude and Redistribution of SECA Modulation

We examine SECA modulation under PFN using Spikformer-2-256 on DVS128-Gesture, with $\eta_{\mathrm{PFN}} \in \{0, 0.02, \ldots, 0.10\}$. For each block ℓ, we measure

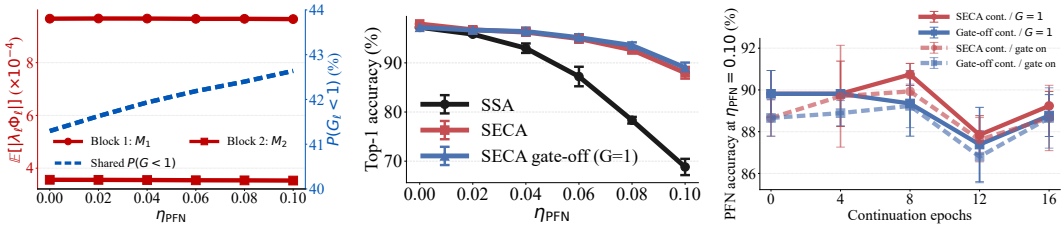
$$M_\ell(\eta) = \mathbb{E}[|\lambda_\ell \Phi_\ell|], \qquad P_{\mathrm{sup}}(\eta) = \Pr(G_\ell < 1). \tag{13}$$

Here, P_{sup} is shared across blocks since $\lambda_\ell > 0$ and $G_\ell < 1$ iff $\rho_i \rho_j < 0$. Statistics cover the full test set and five paired corruption realizations for nonzero severities; learned parameters are reported in Table 7. As shown in Fig. 6a, from clean input to $\eta_{\mathrm{PFN}} = 0.10$, M_1 changes from 9.667×10^{-4} to 9.651×10^{-4} ($\sim 0.17\%$), M_2 from 3.560×10^{-4} to 3.530×10^{-4} ($\sim 0.84\%$), and P_{sup} from 41.29% to 42.64%. Thus, stronger PFN mainly redistributes suppressive interactions rather than increasing average modulation, suggesting that robustness is not obtained through progressively stronger gating but through structured interaction adjustment, motivating the following gate analysis.



(a) t-SNE (b) UMAP

Figure 5: Representation under strong BAN; red boxes indicate overlap.



(a) Modulation statistics. (b) Gate removal. (c) Training-path intervention.

Figure 6: SECA modulation and gate analyses on DVS128-Gesture.

4.6 Inference-Time Gate Removal and Training-Path Effects

We first remove the SECA gate at inference by setting $G \equiv 1$ without retraining. This intervention isolates whether the observed robustness mainly comes from the direct modulation applied during inference or from changes already encoded in the learned parameters. At $\eta_{\mathrm{PFN}} = 0.10$, SSA, full SECA, and gate-off SECA achieve $68.819 \pm 1.654\%$, $87.986 \pm 1.193\%$, and $89.028 \pm 1.030\%$, respectively, over five paired corruption realizations (Fig. 6b). Despite a clean-accuracy change from 97.917% to 97.222%, PFN robustness is largely retained, suggesting that instantaneous inference-time modulation alone does not explain the gain.

We further continue training from the same epoch-197 SECA checkpoint along two matched paths with frozen λ_{raw} and α_s: one retains $G = 1 + \lambda\Phi$, while the other enforces $G \equiv 1$. Under the same gate-off inference operator, the SECA-on path is numerically higher at later checkpoints, reaching $89.24 \pm 0.69\%$ versus $88.77 \pm 1.00\%$ at epoch 16 (Fig. 6c). This comparison removes the effect of different inference operators and focuses on whether structured interaction during continuation training affects the learned representation. The error bars reflect three paired corruption realizations rather than independent training runs; together with Sec. 4.4, this provides descriptive evidence that structured interaction can influence the learned solution, without establishing its full causal origin. Further protocol details are given in Appendix E.4.

5 Conclusion

Prior work on robustness in event-based perception has primarily focused on perturbations at the network input. Here, we advance a complementary perspective by examining the mechanisms governing token–token interactions within the model. Because event tokens preserve the spatial organization and polarity information inherent to the event-generation process, we show that enforcing sensing-derived structural compatibility at the interaction level can improve robustness. Specifically, we demonstrate that a bounded modulation parameterized by only two scalars can induce robustness from clean-only training, with particularly pronounced gains under polarity-corruption scenarios and improvements across seven of nine architecture–corruption settings. While evaluated on three Spiking Transformer families and several event corruptions, broader generalization remains open. SECA constitutes one minimal instantiation of this principle, while more expressive compatibility modeling remains future work.

AI USE STATEMENT

We used generative AI tools to assist with manuscript polishing, including improving clarity, grammar, organization, and readability, and to support the retrieval and discovery of relevant literature. Generative AI tools were also used to assist with limited editing and debugging of research code. All AI-assisted text and code modifications were reviewed and verified by the authors, and cited references were checked against their original sources. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide implementation details, model configurations, training and evaluation protocols, corruption definitions, random seeds, and the experimental environment in Appendix D. Additional SECA implementation details and properties are provided in Appendix C. An anonymized implementation is included in the supplementary material to facilitate reproduction of the reported results.

REFERENCES

Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron Flickner, and Dharmendra Modha. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7388–7397, July 2017. doi: 10.1109/CVPR.2017.781.

Yannick Burkhardt, Simon Schaefer, and Stefan Leutenegger. Superevent: Cross-modal learning of event-based keypoint detection for slam. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8918–8928, October 2025.

Meiqi Cao, Xiangbo Shu, Xin Jiang, Rui Yan, Yazhou Yao, and Jinhui Tang. Exploiting frequency dynamics for enhanced multimodal event-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5969–5979, October 2025.

Meiqi Cao, Jiachao Zhang, Xin Jiang, Rui Yan, Yazhou Yao, Zechao Li, and Xiangbo Shu. Seeing motion through polarity for event-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37075–37085, June 2026.

Jinze Chen, Wei Zhai, Yang Cao, Bin Li, and Zheng-Jun Zha. Event stream filtering via probability flux estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8023–8032, June 2026.

Yuxing Duan. Led: A large-scale real-world paired dataset for event camera denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25637–25647, June 2024.

Rui Fan, Weidong Hao, Juntao Guan, Lai Rui, Tong Wu, Fanhong Zeng, and Lin Gu. SMV-EAR: Bring spatiotemporal multi-view representation learning into efficient event-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6043–6053, June 2026.

Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. doi: 10.1109/TPAMI.2020.3008413.

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 (ICCV)*, pp. 5632–5642, October 2019. doi: 10.1109/ICCV.2019.00573.

Shasha Guo and Tobi Delbruck. Low cost and latency event camera background activity denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):785–795, 2023. doi: 10.1109/TPAMI.2022.3152999.

Yufei Guo, Xiaode Liu, Yuanpei Chen, Weihang Peng, Yuhan Zhang, and Zhe Ma. Spiking transformer: Introducing accurate addition-only spiking self-attention for transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24398–24408, June 2025.

Friedhelm Hamann, Daniel Gehrig, Filbert Febryanto, Kostas Daniilidis, and Guillermo Gallego. Etap: Event-based tracking of any point. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27186–27196, June 2025.

Yuhwan Jeong, Hoonhee Cho, and Kuk-Jin Yoon. Towards robust event-based networks for nighttime via unpaired day-to-night event translation. In *Computer Vision – ECCV 2024*, volume 15125 of *Lecture Notes in Computer Science*, pp. 286–306. Springer, 2024. doi: 10.1007/978-3-031-72855-6_17.

Bin Jiang, Bo Xiong, Bohan Qu, M. Salman Asif, You Zhou, and Zhan Ma. Edformer: Transformer-based event denoising across varied noise levels. In *Computer Vision – ECCV 2024*, volume 15085 of *Lecture Notes in Computer Science*, pp. 200–216. Springer, 2024. doi: 10.1007/978-3-031-73383-3_12.

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 (ICCV)*, pp. 2126–2136, October 2021. doi: 10.1109/ICCV48922.2021.00215.

Lingdong Kong, Dongyue Lu, Xiang Xu, Lai Xing Ng, Wei Tsang Ooi, and Benoit R. Cottereau. Eventfly: Event camera perception from ground to the sky. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1472–1484, June 2025.

Donghyun Lee, Yuhang Li, Youngeun Kim, Shiting Xiao, and Priyadarshini Panda. Spiking transformer with spatial-temporal attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13948–13958, June 2025.

Dianze Li, Jianing Li, Xu Liu, Xiaopeng Fan, and Yonghong Tian. Asynchronous collaborative graph representation for frames and events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1655–1666, June 2025.

Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. CIFAR10-DVS: An event-stream dataset for object classification. *Frontiers in Neuroscience*, 11:309, 2017. doi: 10.3389/fnins.2017.00309.

Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, February 2008. doi: 10.1109/JSSC.2007.914337.

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 (CVPR)*, pp. 25648–25658, June 2024.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861.

Yansong Peng, Yueyi Zhang, Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. GET: Group event transformer for event-based vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6015–6025, October 2023. doi: 10.1109/ICCV51070.2023.00555.

Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. AEGNN: Asynchronous event-based graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12361–12371, June 2022. doi: 10.1109/CVPR52688.2022.01205.

Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Simultaneous motion and noise estimation with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6959–6969, October 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 (CVPR)*, pp. 1731–1740, June 2018. doi: 10.1109/CVPR.2018.00186.

Minggui Teng, Chu Zhou, Hanyue Lou, and Boxin Shi. Nest: Neural event stack for event-based image enhancement. In *Computer Vision – ECCV 2022*, pp. 660–676. Springer, 2022. doi: 10.1007/978-3-031-20068-7_38.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Zengyu Wan, Yang Wang, Ganchao Tan, Yang Cao, and Zheng-Jun Zha. S2n: Suppression-strengthen network for event-based recognition under variant illuminations. In *Computer Vision – ECCV 2022*, volume 13663 of *Lecture Notes in Computer Science*, pp. 716–733. Springer, 2022. doi: 10.1007/978-3-031-20062-5_41.

Shuai Wang, Malu Zhang, Jingya Wang, Dehao Zhang, Yimeng Shan, Jieyuan Zhang, Yichen Xiao, Honglin Cao, Haonan Zhang, Zeyu Ma, Yang Yang, and Haizhou Li. Bipolar self-attention for spiking transformers. In *Advances in Neural Information Processing Systems*, volume 38, pp. 112656–112681, 2025. doi: 10.52202/085713-3401.

Xiao Wang, Zongzhen Wu, Bo Jiang, Zhimin Bao, Lin Zhu, Guoqi Li, Yaowei Wang, and Yonghong Tian. HARDVS: Revisiting human activity recognition with dynamic vision sensors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5615–5623, 2024. doi: 10.1609/aaai.v38i6.28372.

Wenjie Wei, Xiaolong Zhou, Malu Zhang, Ammar Belatreche, Qian Sun, Yimeng Shan, Dehao Zhang, Zijian Zhou, Zeyu Ma, Yang Yang, and Haizhou Li. TP-spikformer: Token pruned spiking transformer. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=L5llQD0nMf.

Sheng Wu, Hang Sheng, Hui Feng, and Bo Hu. Egsst: Event-based graph spatiotemporal sensitive transformer for object detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 120526–120548, 2024. doi: 10.52202/079017-3831.

Senyan Xu, Shuai Chen, Chuanfu Shen, Kean Liu, Zhijing Sun, Chengzhi Cao, and Xueyang Fu. Eventgait: Towards robust gait recognition with event streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22323–22334, June 2026.

Shengdong Xue, Haoxiang Ma, Hao Chen, Zhen Yang, and Yongjian Deng. Event-based motion deblurring using task-oriented 3d gaussian event representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29547–29556, June 2026.

Man Yao, Jiakui Hu, Zhaokun Zhou, Li Yuan, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64043–64058, 2023. doi: 10.52202/075280-2798.

Man Yao, Jiakui Hu, Tianxiang Hu, Yifan Xu, Zhaokun Zhou, Yonghong Tian, Bo Xu, and Guoqi Li. Spike-driven transformer V2: Meta spiking neural network architecture inspiring the design of next-generation neuromorphic chips. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1SIBN5Xyw7.

Man Yao, Xuerui Qiu, Tianxiang Hu, Jiakui Hu, Yuhong Chou, Keyu Tian, Jianxing Liao, Luziwei Leng, Bo Xu, and Guoqi Li. Scaling spike-driven transformer with efficient spike firing approximation training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 2973–2990, 2025. doi: 10.1109/TPAMI.2025.3530246.

Zishu Yao, Xiang-Xiang Su, Shengning Zhou, Guang-Yong Chen, Guodong Fan, and Xing Chen. Bievlight: Bi-level learning of task-aware event refinement for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41541–41550, June 2026.

Dehao Zhang, Fukai Guo, Shuai Wang, Jingya Wang, Jieyuan Zhang, Yimeng Shan, Malu Zhang, Yang Yang, and Haizhou Li. Neural dynamics self-attention for spiking transformers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=jJedqisfOt.

Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. Spiking transformers for event-based single object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8791–8800, June 2022. doi: 10.1109/CVPR52688.2022.00860.

Chenlin Zhou, Han Zhang, Zhaokun Zhou, Liutao Yu, Liwei Huang, Xiaopeng Fan, Li Yuan, Zhengyu Ma, Huihui Zhou, and Yonghong Tian. Qkformer: Hierarchical spiking transformer using q-k attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 13074–13098, 2024a. doi: 10.52202/079017-0416.

Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Eventbind: Learning a unified representation to bind them all for event-based open-world understanding. In *Computer Vision – ECCV 2024*, pp. 477–494. Springer, 2024b. doi: 10.1007/978-3-031-72897-6_27.

Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Li Yuan. Spikformer: When spiking neural network meets transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=frE4fUwz_h.

Zhaokun Zhou, Yijie Lu, Yanhao Jia, Kaiwei Che, Jun Niu, Liwei Huang, Xinyu Shi, Yuesheng Zhu, Guoqi Li, Zhaofei Yu, and Li Yuan. Spiking transformer with experts mixture. In *Advances in Neural Information Processing Systems*, volume 37, pp. 10036–10059, 2024c. doi: 10.52202/079017-0322.

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 (ICCV)*, pp. 12800–12810, October 2023. doi: 10.1109/ICCV51070.2023.01180.

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 (CVPR)*, pp. 5819–5828, June 2024.

APPENDIX

A TOKEN-ALIGNED POLARITY REPRESENTATION

The input event stream is first converted into T event frames using the same preprocessing as the feature branch, with events partitioned by event count rather than by fixed temporal duration. Let $\mathbf{F} \in \mathbb{R}^{T\times B\times 2\times H\times W}$ denote the resulting two-channel event-frame tensor. For temporal step t, batch element b, and token index n, we average each polarity channel over the spatial region $\mathcal{R}_n$:

$$a_{t,b,n}^{(c)} = \frac{1}{|\mathcal{R}_n|} \sum_{(u,v)\in\mathcal{R}_n} F_{t,b,c,u,v}, \qquad c \in \{0,1\}. \tag{14}$$

The token-aligned polarity descriptor is then defined by

$$\rho_{t,b,n} = \mathcal{A}\Big(a_{t,b,n}^{(0)}, a_{t,b,n}^{(1)}\Big) = \frac{a_{t,b,n}^{(0)} - a_{t,b,n}^{(1)}}{a_{t,b,n}^{(0)} + a_{t,b,n}^{(1)} + \epsilon}, \qquad \epsilon = 10^{-6}. \tag{15}$$

For the DVS128-Gesture configuration, $\mathcal{R}_n$ is a non-overlapping 16×16 pooling region, producing an 8×8 descriptor grid. Hence $\rho_{t,b,n} \in [-1,1]$, with its sign indicating which polarity channel dominates locally and its magnitude reflecting the channel imbalance. If both pooled channels are zero, then $\rho_{t,b,n} = 0$, corresponding to an inactive token region.

The descriptor grid is aligned with the feature-token grid at the same temporal and token indices, although the two branches need not share identical pixel-level receptive fields. SECA uses the polarity descriptors through

$$P_{t,b,i,j} = \rho_{t,b,i}\rho_{t,b,j}, \tag{16}$$

which forms the polarity component of the full prior $\Phi_{t,b,i,j} = P_{t,b,i,j}\kappa_{ij}$. Positive and negative values encode same- and opposite-sign channel tendencies, respectively, and are both valid structural cues. For an empty token, $\rho_{t,b,n} = 0$, so any associated prior entry satisfies $\Phi_{t,b,i,j} = 0$ and therefore $G_{t,b,i,j} = 1$, introducing no structural modulation.

B EVENT-GENERATION MOTIVATION

Section 3.2 uses the event-triggering mechanism only to motivate polarity as a sensing-derived structural cue. Here we make its connection to local image motion more explicit. Let $L(x, y, \tau)$ denote log intensity. Under the local brightness-constancy assumption, a moving image pattern approximately satisfies the following constraint:

$$\frac{\partial L}{\partial \tau} + \nabla L^\top \mathbf{v} \approx 0, \tag{17}$$

where $\mathbf{v}$ denotes the local image velocity. Over a short interval $\Delta\tau$, this gives

$$\Delta L \approx -\nabla L^\top \mathbf{v}\, \Delta\tau. \tag{18}$$

Combining this relation with the threshold-crossing model in Sec. 3.2 suggests

$$p_m \approx \operatorname{sign}(\Delta L_m) \approx \operatorname{sign}\left(-\nabla L^\top \mathbf{v}\right), \tag{19}$$

showing that event polarity is related to the interaction between local image gradients and motion rather than being an arbitrary binary label.

Spatial proximity provides complementary information. Within a locally coherent motion field, neighboring event responses are typically organized by nearby image structure and related motion, motivating a distance-dependent locality cue. However, proximity does not imply identical polarity or common physical origin: gradient orientation may change across edges, both ON and OFF responses can occur within the same local structure, and multiple motions may coexist within a neighborhood. Therefore, neither spatial proximity nor polarity agreement is treated as a deterministic consistency rule, but rather as a soft structural prior.

This interpretation motivates using spatial locality and polarity jointly as soft compatibility cues in equation 5. The resulting prior should not be interpreted as a generative likelihood or as a hard physical constraint; rather, it encodes sensing-derived structure that can guide, but not determine, token interactions under diverse event corruption conditions.

C SECA PROPERTIES AND IMPLEMENTATION

Proof of the bounded modulation properties. We provide the proof of the properties stated in Sec. 3.3. From $\rho_{t,b,n} \in [-1, 1]$ and $\kappa_{ij} \in (0, 1]$, the prior in equation 5 satisfies

$$|\Phi_{t,b,i,j}| = |\rho_{t,b,i}\rho_{t,b,j}|\kappa_{ij} \leq 1. \tag{20}$$

Since $\lambda = \frac{1}{2}\,\mathrm{Sigmoid}(\lambda_{\mathrm{raw}}) \in (0, 0.5)$,

$$-\lambda \leq \lambda\Phi_{t,b,i,j} \leq \lambda, \tag{21}$$

which directly yields the gate bound in equation 9. Because the resulting gate is strictly positive, multiplication by $G_{t,b,i,j}$ preserves the sign of every nonzero SSA correlation, and

$$\frac{|S^{\mathrm{SECA}}_{t,b,h,i,j}|}{|S^{\mathrm{SSA}}_{t,b,h,i,j}|} = G_{t,b,i,j} \in (0.5, 1.5) \tag{22}$$

whenever $S^{\mathrm{SSA}}_{t,b,h,i,j} \neq 0$. The multiplicative form also immediately implies $S^{\mathrm{SSA}}_{t,b,h,i,j} = 0 \Rightarrow S^{\mathrm{SECA}}_{t,b,h,i,j} = 0$. Finally, SECA continuously recovers the unmodulated interaction as $\lambda \to 0$:

$$\mathbf{S}^{\mathrm{SECA}} \to \mathbf{S}^{\mathrm{SSA}}. \tag{23}$$

Initialization. We initialize $\lambda_{\mathrm{raw}} = -2.197$, giving $\mathrm{Sigmoid}(\lambda_{\mathrm{raw}}) \approx 0.1$ and hence $\lambda \approx 0.05$. The spatial-scale parameter is initialized as $\alpha_s = 0$, corresponding to $\sigma_s^2 = 1$. Thus, optimization starts with a weak structural modulation while allowing both scalars to adapt jointly with the backbone from a stable initial regime.

Tensor-level implementation. For a fixed token grid, we precompute the squared-distance matrix

$$\mathbf{D} = [d_{ij}^2]_{i,j=1}^{N} \in \mathbb{R}^{N\times N}. \tag{24}$$

At each forward pass, the spatial kernel and event-derived prior are evaluated as

$$\boldsymbol{\kappa} = \exp\left(-\frac{\mathbf{D}}{\exp(\alpha_s)}\right), \qquad \boldsymbol{\Phi}_{t,b} = \left(\boldsymbol{\rho}_{t,b}\boldsymbol{\rho}_{t,b}^{\top}\right) \odot \boldsymbol{\kappa}. \tag{25}$$

The resulting gate $\mathbf{G}_{t,b} \in \mathbb{R}^{N\times N}$ is inserted along the head dimension and broadcast to all heads,

$$\mathbf{S}^{\mathrm{SECA}}_{t,b,h} = \mathbf{S}^{\mathrm{SSA}}_{t,b,h} \odot \mathbf{G}_{t,b}. \tag{26}$$

No head-wise copy of the prior is required conceptually, and the implementation can exploit standard broadcasting over the head dimension. The cached geometry is independent of the input, whereas $\boldsymbol{\rho}$, $\boldsymbol{\Phi}$, and $\mathbf{G}$ are recomputed for each input and temporal step.

D EXPERIMENTAL DETAILS

Datasets. We evaluate SECA on DVS128-Gesture (Amir et al., 2017), CIFAR10-DVS (Li et al., 2017), N-Cars (Sironi et al., 2018), and HARDVS (Wang et al., 2024). DVS128-Gesture is the primary benchmark for robustness and mechanism analyses, while the remaining datasets evaluate cross-dataset and real-world applicability across diverse event-based recognition scenarios. We use the standard train/test partitions and report Top-1 classification accuracy.

Backbones. We evaluate Spikformer-2-256 (Zhou et al., 2023), SDT-8-384 and SDT-8-512 (Yao et al., 2023), and SDT-V3-2-360 and SDT-V3-8-360 (Yao et al., 2025). The suffixes denote network depth and embedding dimension. Baselines retain their native attention operators (SSA or SDSA), while SECA replaces only the corresponding interaction operator; all other backbone settings are kept unchanged across architectures and model scales.

Training and comparison protocol. For each backbone–dataset pair, baseline and SECA models use identical preprocessing, temporal discretization, optimization, augmentation, and evaluation settings. Unless otherwise stated, $T = 16$ temporal steps are used; Sec. E.5 varies $T \in \{4, 8, 16, 32, 64\}$. Robustness models are trained only on clean data, with BAN, PFN, and HPN applied exclusively at test time. The default random seed is 2026; experiments explicitly evaluating seed variation use $\{2026, 2027, 2028, 2029, 2030\}$.

Implementation environment. All experiments are conducted on a workstation equipped with two NVIDIA L20 GPUs, each with 46 GB of memory. The software environment consists of Python 3.10.20, PyTorch 2.5.1 with CUDA 12.1, and SpikingJelly 0.0.0.0.14.

Event corruption protocol. Corruptions are applied to the discretized event representation

$$\mathbf{X}_{\mathrm{evt}} \in \mathbb{R}^{T\times 2\times H\times W}, \tag{27}$$

where the two channels represent the two polarity channels. For SECA, the token-aligned polarity descriptor and prior are recomputed from the corrupted input:

$$\mathbf{X}_{\mathrm{evt}} \to \mathrm{Corrupt}(\mathbf{X}_{\mathrm{evt}}) \to \boldsymbol{\rho} \to \boldsymbol{\Phi} \to \mathrm{SECA}\,. \tag{28}$$

Background Activity Noise (BAN). BAN injects background activity at random temporal steps, spatial locations, and polarity channels while preserving the original input, simulating spurious sensor-triggered events commonly observed in event streams. Its severity is controlled by ρ_{BAN}.

Polarity Flip Noise (PFN). PFN flips the polarity assignment of a fraction η_{PFN} of existing activity while preserving its spatial–temporal bins and event count.

Hot-Pixel Noise (HPN). HPN selects fixed spatial locations and injects persistent spurious activity across temporal steps. Its severity is controlled by the hot-pixel density d_{HPN}.

For the reported robustness summaries, the severity grids are $\rho_{\mathrm{BAN}} \in \{0.20, 0.40, \ldots, 2.00\}$, $\eta_{\mathrm{PFN}} \in \{0.01, 0.02, \ldots, 0.10\}$, and $d_{\mathrm{HPN}} \in \{0.02, 0.04, \ldots, 0.20\}$. Mean ΔAcc averages the clean-relative degradation over all ten nonzero severities, while Max-Severity ΔAcc reports the degradation at the largest reported severity. The fixed-severity analyses use $\rho_{\mathrm{BAN}} = 2.0$, $\eta_{\mathrm{PFN}} = 0.10$, and $d_{\mathrm{HPN}} = 0.10$.

Evaluation metrics. For corruption severity s, we measure clean-relative degradation as

$$\Delta\mathrm{Acc}(s) = \mathrm{Acc}_{\mathrm{clean}} - \mathrm{Acc}(s). \tag{29}$$

Mean ΔAcc averages over all tested non-zero severities, while Max ΔAcc reports the degradation at the strongest tested severity. Both are measured in percentage points (p.p.); lower is better. Unless otherwise stated, degradation is computed relative to the clean accuracy of the same trained model.

E Additional Experimental Results and Analyses

E.1 Cross-Backbone and Cross-Dataset Robustness

To examine whether the robustness behavior in Sec. 4.1 is specific to Spikformer, we further evaluate SECA with SDT-8-384 and SDT-V3-2-360 on DVS128-Gesture. All models follow the same clean-only training and test-time corruption protocol described in Appendix D, with the corresponding SDSA model used as the baseline. Figure 7 reports the complete accuracy curves under BAN, PFN, and HPN. Across the two additional backbones, SECA reduces corruption degradation in most settings, although the effect is architecture- and corruption-dependent. In particular, SDT-8-384 benefits under BAN and PFN and remains nearly unchanged in mean HPN degradation, whereas SDT-V3-2-360 improves under BAN and HPN but exhibits larger clean-relative degradation under PFN. These results show that the robustness behavior extends beyond the Spikformer backbone, while also revealing backbone-specific differences.

As an additional cross-dataset evaluation, we test SDT-8-384 on CIFAR10-DVS using the same clean-only training and test-time corruption protocol. The clean accuracies of the Base and SECA variants are 75.0% and 76.4%, respectively. Table 6 reports the clean-relative degradation over three paired corruption realizations. SECA reduces both mean and maximum-severity degradation under BAN and HPN, but exhibits larger clean-relative degradation under PFN. Thus, the robustness benefit transfers to a second event dataset for some corruption types, but does not hold uniformly across corruptions, suggesting that the effect depends on both event distribution and corruption characteristics rather than being universally dominant.

Table 6: Cross-dataset robustness of SDT-8-384 on CIFAR10-DVS. B/S denotes Base/SECA; ΔAcc is clean-relative and reported as mean ± std. over three paired corruption realizations.

Noise	Mean ΔAcc B/S ↓	Max ΔAcc B/S ↓
BAN	36.85 ± 0.13 / $\mathbf{35.05 \pm 0.06}$	39.20 ± 0.00 / $\mathbf{36.80 \pm 0.00}$
PFN	$\mathbf{4.99 \pm 0.44}$ / 6.14 ± 0.21	$\mathbf{5.93 \pm 0.15}$ / 7.80 ± 0.70
HPN	30.95 ± 0.37 / $\mathbf{29.66 \pm 0.15}$	43.53 ± 0.98 / $\mathbf{41.67 \pm 0.31}$

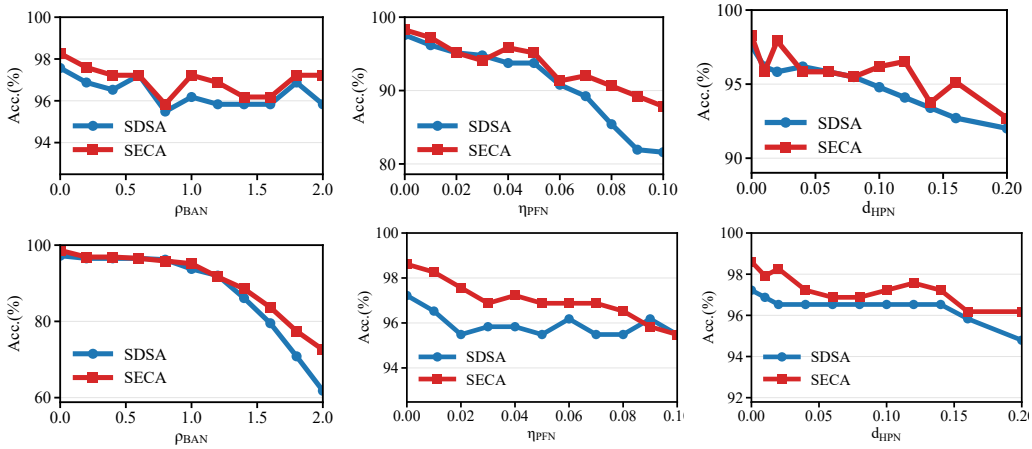


Figure 7: Additional robustness curves on DVS128-Gesture. Top: SDT-8-384; bottom: SDT-V3-2-360. Columns correspond to BAN, PFN, and HPN.

E.2 Additional Representation Visualizations

We further examine the learned representations under clean inputs, PFN, and HPN using both t-SNE and UMAP, complementing the BAN analysis in Sec. 4.3. Using two projection methods helps assess whether the qualitative pattern is specific to a single embedding procedure. As in the main text, each embedding is fitted independently; therefore, absolute distances, cluster sizes, and relative positions across different panels should not be compared directly. We instead focus on qualitative properties within each projection, including class compactness, separation, and cross-class overlap.

As shown in Fig. 8, SSA and SECA exhibit broadly similar class organization on clean inputs, consistent with their comparable clean recognition performance. Under PFN and HPN, however, SSA representations become more dispersed and exhibit more pronounced local overlap between classes, whereas SECA tends to preserve visually tighter and better-separated class structure across both t-SNE and UMAP. The effect is qualitative rather than a direct measure of representation geometry, but the similar tendency across the two projection methods provides complementary support for the corruption robustness reported in Sec. 4.1.

E.3 Block-Wise SECA Parameters

Table 7 reports the learned block-wise SECA parameters of Spikformer-2-256 on DVS128-Gesture, where $\lambda_\ell = 0.5\,\mathrm{Sigmoid}(\lambda_\ell^{\mathrm{raw}})$ controls the modulation strength and $\sigma_{s,\ell}^2 = \exp(\alpha_{s,\ell})$ controls the spatial decay of the prior. Starting from $\lambda_\ell \approx 0.05$ and $\sigma_{s,\ell}^2 = 1$, both blocks learn smaller spatial scales, while the second block also learns a substantially smaller modulation coefficient. Specifically, Block 1 retains $\lambda_1 = 0.04448$ with $\sigma_{s,1}^2 = 0.58260$, whereas Block 2 learns $\lambda_2 = 0.02079$ and a more localized spatial scale of $\sigma_{s,2}^2 = 0.36456$. This difference is consistent with the larger modulation magnitude observed for Block 1 in Fig. 6a, although the realized modulation also depends on the input-dependent prior $\mathbf{\Phi}_\ell$.

Table 7: Learned block-wise SECA parameters on DVS128-Gesture.

Block	$\boldsymbol{\lambda_\ell}$	$\boldsymbol{\alpha_{s,\ell}}$	$\boldsymbol{\sigma_{s,\ell}^2}$
1	0.04448	−0.54026	0.58260
2	0.02079	−1.00907	0.36456

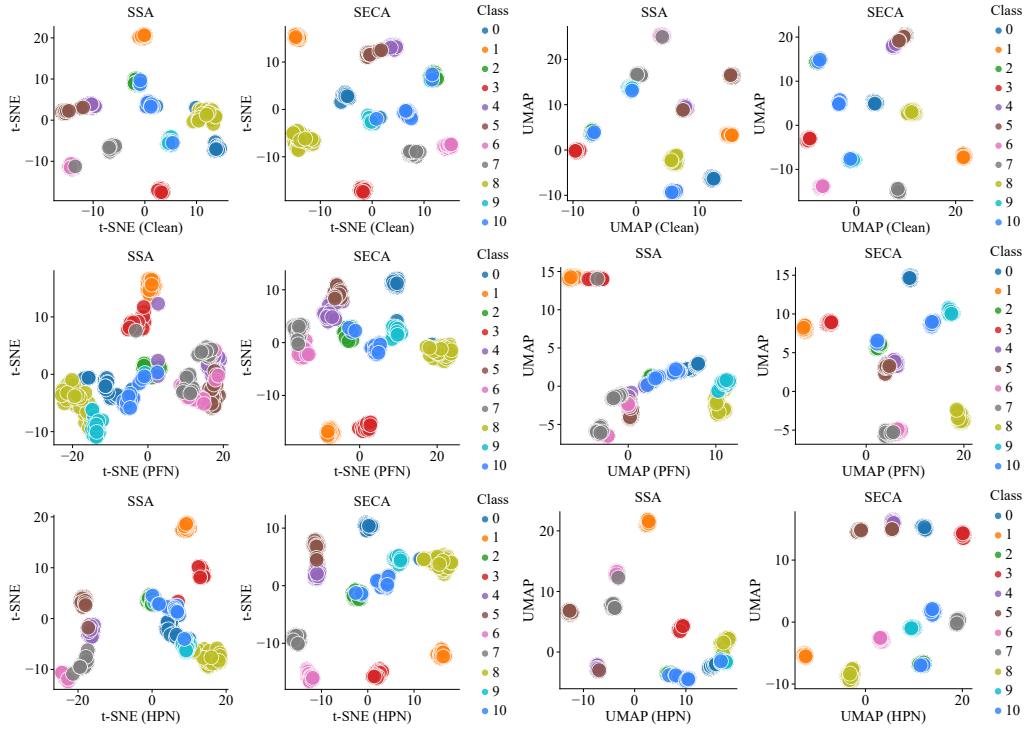


Figure 8: Additional DVS128-Gesture representations: clean, PFN, and HPN (top to bottom), with t-SNE and UMAP (left to right).

Table 8: Inference-time gate removal on DVS128-Gesture at $\eta_{\mathrm{PFN}} = 0.10$.

Model	**Clean Acc. (%)**	**PFN Acc. (%)**	**ΔAcc (p.p.) ↓**
SSA	97.222	68.819 ± 1.654	28.403 ± 1.654
Full SECA	97.917	87.986 ± 1.193	9.931 ± 1.193
SECA gate-off ($G \equiv 1$)	97.222	89.028 ± 1.030	8.194 ± 1.030

All learned scalars remain fixed during corruption evaluation; only the token-aligned polarity descriptors and the resulting $\mathbf{\Phi}_\ell$ are recomputed from each corrupted input. Therefore, the severity-dependent changes in Fig. 6a reflect changes in event-derived structure rather than test-time adaptation of λ_ℓ or $\sigma^2_{s,\ell}$. The learned values also indicate that SECA operates with relatively small block-wise modulation coefficients, consistent with the small average gate deviations reported in the main text and further supporting its bounded correction design.

E.4 Gate Removal and Training-Path Intervention

Table 8 reports the exact inference-time gate-removal results corresponding to Sec. 4.6. SECA gate-off uses the same trained checkpoint as full SECA, with only the multiplicative gates disabled by setting $G \equiv 1$; no parameters are retrained or replaced. This intervention separates robustness obtained from instantaneous inference-time modulation from robustness already encoded in the learned model parameters. PFN results are reported as mean $\pm$ standard deviation over five paired corruption realizations, with ΔAcc computed relative to each model's own clean accuracy. For the continuation intervention, both paths start from the same epoch-197 SECA checkpoint with identical optimizer and scheduler states, while all λ_{raw} and α_s parameters are frozen. One path retains $G = 1 + \lambda\Phi$ during continuation training, whereas the other enforces $G \equiv 1$, with all remaining conditions matched. Both models are evaluated with the gate enabled and disabled at each checkpoint, forming the 2×2 intervention in Fig. 6c. Comparing the two paths under the same gate-off inference operator isolates differences accumulated during continuation training rather than differences caused

Table 10: Clean and corruption performance on HARDVS under BAN, PFN, and HPN. Lower ΔAcc is better; Δ denotes SECA−SDSA.

Setting	Metric	SDSA	SECA	Δ
Clean	Acc. (%)	**52.21**	51.88	-0.33
BAN	Acc. (%) @ $\rho_{\mathrm{BAN}} = 1.0$	40.86	**42.04**	**+1.18**
	Mean ΔAcc (p.p.) ↓	5.80	**5.19**	**-0.61**
	Max ΔAcc (p.p.) ↓	11.35	**9.84**	**-1.51**
PFN	Acc. (%) @ $\eta_{\mathrm{PFN}} = 0.10$	**48.01**	47.86	-0.15
	Mean ΔAcc (p.p.) ↓	1.81	**1.57**	**-0.24**
	Max ΔAcc (p.p.) ↓	4.20	**4.02**	**-0.18**
HPN	Acc. (%) @ $d_{\mathrm{HPN}} = 0.10$	46.05	**46.21**	**+0.16**
	Mean ΔAcc (p.p.) ↓	3.86	**3.81**	**-0.05**
	Max ΔAcc (p.p.) ↓	6.16	**5.67**	**-0.49**

by inference-time processing. The error bars reflect three paired PFN corruption realizations rather than independent continuation-training runs, so the results are interpreted as descriptive evidence of a training-path effect rather than an estimate of training-run variability. Because continuation begins from an already trained SECA checkpoint, this intervention does not identify the causal origin of robustness acquired during earlier training.

E.5 Temporal-Resolution Compatibility

Using SDT-V3-2-360 on DVS128-Gesture, we compare SDSA and SECA at $T \in \{4, 8, 16, 32, 64\}$ under matched settings to examine compatibility across temporal resolutions. As shown in Table 9, the absolute accuracy difference between the two methods never exceeds 0.694 p.p. over the full range of T: SECA is slightly lower at $T = 4$ and $T = 8$, improves over SDSA by 0.347 p.p. at $T = 16$, and matches SDSA at $T = 32$ and $T = 64$. The absence of systematic degradation as T increases suggests that SECA does not rely on a particular temporal discretization to preserve clean-data performance. This is consistent with its design, since the spatial–polarity prior is constructed independently at each temporal step and modulates token interactions only within that step, without introducing additional cross-step interaction or temporal state. Consequently, changing T affects the backbone's temporal resolution while leaving the form of the SECA operation unchanged. These results support the temporal-resolution compatibility of SECA across a broad range of timesteps, although this experiment evaluates clean recognition only and does not assess corruption robustness across different T.

Table 9: Clean Top-1 accuracy (%) across temporal resolutions on DVS128-Gesture. Δ denotes SECA−SDSA (p.p.).

Timesteps	SDSA	SECA	Δ (p.p.)
4	**95.833**	95.139	-0.694
8	**97.917**	97.569	-0.348
16	97.917	**98.264**	**+0.347**
32	**97.569**	**97.569**	0.000
64	**97.222**	**97.222**	0.000

E.6 Real-World Evaluation on HARDVS

We further evaluate SECA on HARDVS, a large-scale real-world event-based human activity recognition benchmark (Wang et al., 2024), using SDT-V3-2-360 to examine whether the robustness trends extend beyond DVS128-Gesture. As shown in Table 10, SECA achieves 51.88% clean accuracy compared with 52.21% for SDSA, a difference of only 0.33 p.p. Across corruption sweeps, SECA yields lower clean-relative degradation for all three corruptions: mean/max ΔAcc decrease from $5.80/11.35$ to $5.19/9.84$ p.p. under BAN, from $1.81/4.20$ to $1.57/4.02$ p.p. under PFN, and from $3.86/6.16$ to $3.81/5.67$ p.p. under HPN. The gains are most evident under BAN and smaller for PFN and HPN. At $\eta_{\mathrm{PFN}} = 0.10$, SECA attains 47.86% absolute accuracy versus 48.01% for SDSA, so the PFN result reflects lower clean-relative degradation rather than an absolute-accuracy advantage. Overall, although the gains are smaller than on DVS128-Gesture, SECA remains compatible with a larger and more diverse real-world event-recognition setting.