MOTION BEFORE EVENTS: IMU-GUIDED OPTICAL CONTROL AND TIMESTAMP-AWARE COMPENSATION

Anonymous authors
Paper under double-blind review

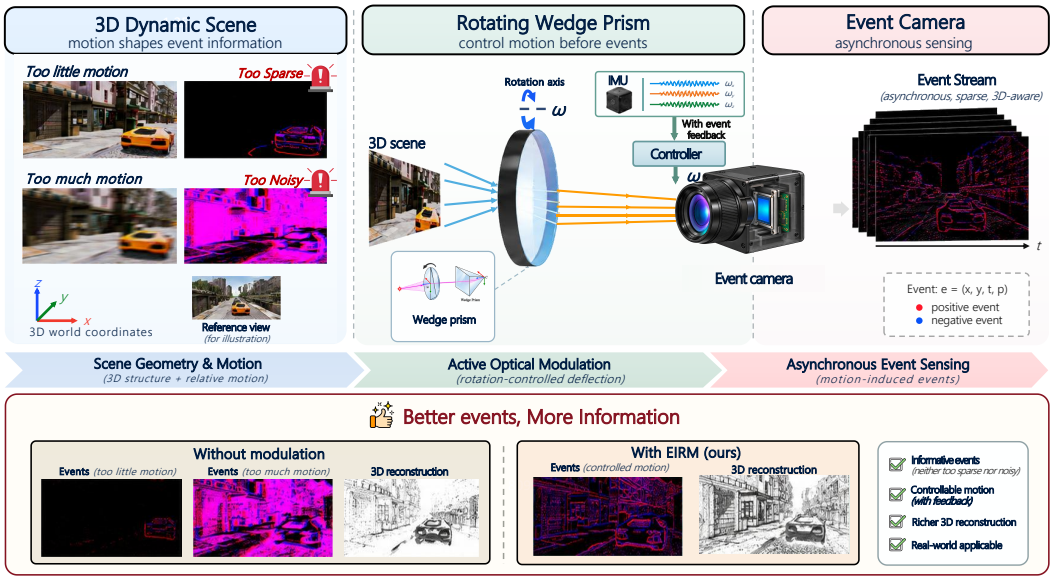


Figure 1: **Motion-adaptive event sensing with EIRM.** EIRM regulates optical excitation before event generation to maintain informative observations across changing motion conditions.

ABSTRACT

Neuromorphic devices like event cameras enable low-latency visual sensing, but their information content depends strongly on image motion. Weak motion may leave important structures under-observed, whereas strong motion can produce excessive and redundant events. Inspired by active modulation in biological vision, we present *Event-Inertial Retinal Modulation* (**EIRM**), a bio-inspired framework that adaptively regulates optical excitation according to the current sensing state. EIRM uses a causal temporal policy to select discrete wedge-prism rotation speeds from recent visual, inertial, and event-derived observations, increasing modulation when natural motion is insufficient and reducing it when event excitation is already adequate. A complementary rule-based feedback loop regulates controlled camera motion from aligned-event activity. To assess whether the resulting events are informative rather than merely abundant, we introduce *Target Event Observability* (TEO), which jointly measures structural coverage, event purity, and temporal persistence. We further construct a hybrid benchmark combining controllable synthetic scenes, geometry-consistent reconstructions, and real-world visual–inertial trajectories. Experiments show that EIRM recovers useful structure under weak motion while suppressing unnecessary excitation under strong motion.

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.

1 INTRODUCTION

Unlike conventional RGB cameras, event cameras asynchronously measure per-pixel changes in log intensity. Their high temporal resolution, high dynamic range, low latency, and reduced motion blur make them well suited to high-speed vision under challenging illumination (Lichtsteiner et al., 2008; Gallego et al., 2022; Feng et al., 2025; Low & Lee, 2023; Yu et al., 2026; Wu et al., 2026; Ma et al., 2026). Yet their change-driven sensing mechanism also makes event information intrinsically motion-dependent: weak relative image motion may leave informative structures under-observed, whereas strong motion can produce excessive and redundant firing. Thus, more motion does not necessarily imply more useful event information.

This dependence has motivated active approaches that shape event acquisition through camera motion, gaze shifts, fixational motion, or adaptive sensor parameters (Mishra et al., 2017; Yousefzadeh et al., 2018; Testa et al., 2020; Delbruck et al., 2021). Most closely related to our setting, active optical modulation can introduce image motion without moving the entire camera (He et al., 2024). However, existing sensing motion is typically prescribed independently of the motion and event activity already present. A fixed modulation pattern useful under weak motion may become unnecessary or even counterproductive once sufficient natural motion is available. This raises a central question: *How should optical excitation adapt to changing motion conditions so that event sensing remains informative without becoming insufficient or excessively redundant?*

Biological vision suggests a useful principle. Microscopic eye movements convert spatial luminance variation into temporally structured retinal responses during fixation (Rucci et al., 2007; Kuang et al., 2012), while stabilization mechanisms reduce excessive retinal motion under stronger external motion. Rather than reproducing a specific biological mechanism, we abstract this broader principle: sensory motion should depend on the current motion state, shifting the objective from simply generating more events to maintaining an informative sensing regime across changing motion.

Based on this idea, we present *Event-Inertial Retinal Modulation* (EIRM), a bio-inspired active sensing framework that adaptively regulates the motion presented to an event camera. A causal temporal policy adjusts wedge-prism excitation from recent visual, inertial, and event-derived observations, strengthening optical modulation when natural motion is insufficient and reducing it when additional excitation becomes unnecessary. A complementary event-driven feedback path regulates the controlled camera motion according to aligned-event activity, allowing the system to accelerate, maintain, or decelerate the sensing motion as conditions change. Meanwhile, the known prism trajectory enables timestamp-aware compensation of the induced optical motion, yielding aligned events on a common reference plane for subsequent sensing and feedback. Together, these components form a causal closed-loop acquisition process that regulates motion before events are generated.

Evaluating such behavior requires more than event count: dense streams may contain substantial redundancy, while sparse streams may miss important target structure. We therefore introduce *Target Event Observability* (TEO), a target-centric metric that jointly evaluates structural coverage, event purity, and temporal persistence. We further construct a hybrid benchmark combining controlled synthetic scenes, geometry-consistent rendering from reconstructed real scenes, and natural image motion with real-world RGB–IMU trajectories. All conditions share the same pre-event optical modulation and event-generation pipeline, enabling controlled comparison of fixed and adaptive sensing across diverse motion regimes. Across controlled and realistic motion conditions, EIRM recovers under-observed structure while adapting sensing motion to avoid unnecessary excitation when sufficient motion is already present.

Our contributions are:

- We present EIRM, an IMU-guided motion-adaptive sensing framework that jointly coordinates optical excitation and controlled camera motion before event generation, while using timestamp-aware compensation to preserve coherent event geometry.
- We introduce *Target Event Observability* (TEO), which measures useful target observation beyond event abundance under diverse camera–object relative motion conditions.
- We construct a hybrid simulation-and-real-world benchmark covering diverse camera–object motion regimes and realistic visual–inertial dynamics across sensing conditions.

2 Related Work

2.1 Bio-Inspired and Active Event Sensing

Event cameras respond to temporal changes in log intensity, making their measurements intrinsically motion-dependent. This has motivated active sensing strategies that deliberately introduce motion to reveal otherwise weak or static visual structure. Early work generated events from static images using prescribed saccadic trajectories (Orchard et al., 2015), while subsequent studies exploited active camera motion for segmentation and recognition (Mishra et al., 2017; Yousefzadeh et al., 2018). Inspired by fixational eye movements, Testa et al. (Testa et al., 2020; 2023) further showed that small controlled motions can convert spatial contrast into temporally informative event activity.

Active sensing can also modify acquisition without moving the entire camera. Delbruck et al. (Delbruck et al., 2021) adapt event-camera parameters to regulate event rate and noise, while AMI-EV (He et al., 2024) uses a continuously rotating wedge prism to induce and compensate microsaccade-like image motion. EIRM instead adapts optical excitation to the current sensing state and combines it with event-driven motion feedback, rather than prescribing a fixed optical trajectory. This design is broadly inspired by biological stabilization mechanisms such as the vestibulo-ocular reflex (Angelaki & Hess, 2005), without attempting to reproduce the underlying physiology. Related event–inertial systems mainly use IMU measurements for downstream motion compensation, odometry, or state estimation (Zhu et al., 2017; Mueggler et al., 2018; Rosinol Vidal et al., 2018). In EIRM, by contrast, inertial cues directly influence the acquisition policy before event generation.

2.2 Event Regulation and Observability

Event activity alone is not a sufficient indicator of sensing quality: low activity may indicate missing structure, whereas excessive activity can introduce substantial redundancy. Sensor-level feedback regulates global statistics such as event rate and noise (Delbruck et al., 2021), while motion-aware suppression methods reduce redundant motion-induced events (Pellerito et al., 2026). These approaches mainly regulate the sensor response or the emitted event stream. EIRM instead acts before event generation, adapting optical excitation while aligned-event feedback continuously and causally determines whether controlled camera motion should accelerate, remain unchanged, or decelerate.

Motion-consistency objectives provide another view of event quality. Contrast-maximization methods warp events under a motion model and optimize the sharpness of the resulting image of warped events (Gallego et al., 2018), but geometric sharpness does not directly indicate whether relevant scene structure has been sufficiently observed. Other active-event systems evaluate acquisition through downstream tasks such as recognition, tracking, reconstruction, or state estimation (Yousefzadeh et al., 2018; He et al., 2024). We therefore introduce *Target Event Observability* (TEO), which jointly evaluates target-edge coverage, event purity, and temporal persistence, providing a useful acquisition-level complement to event-rate, motion-consistency, and task-specific metrics.

2.3 Event Simulation and Evaluation Benchmarks

Event simulators provide carefully controlled acquisition conditions that are difficult to reproduce systematically with physical hardware. ESIM combines rendering with adaptive temporal sampling (Rebecq et al., 2018), while VID2E and v2e convert temporally dense frame sequences into event streams (Gehrig et al., 2020; Hu et al., 2021). DVS-Voltmeter further models stochastic sensor behavior (Lin et al., 2022), while reconstructed 3D scenes provide controllable camera trajectories with realistic appearance. GS2E (Li et al., 2025), for example, uses Gaussian-splatting reconstructions to generate geometrically consistent event observations under controlled camera motion.

Real-world datasets provide complementary visual and motion statistics. MVSEC (Zhu et al., 2018) and DSEC (Gehrig et al., 2021) provide synchronized event, frame, inertial, and geometric measurements, but their prerecorded trajectories fix the relationship between camera motion, scene motion, and event excitation. Existing resources therefore offer either controllability or realistic recorded dynamics, but not matched relative-motion conditions for evaluating adaptive acquisition. Our benchmark closes this gap by combining controllable camera–object motion with realistic appearance and inertial dynamics, enabling consistent comparison between fixed and adaptive optical sensing.

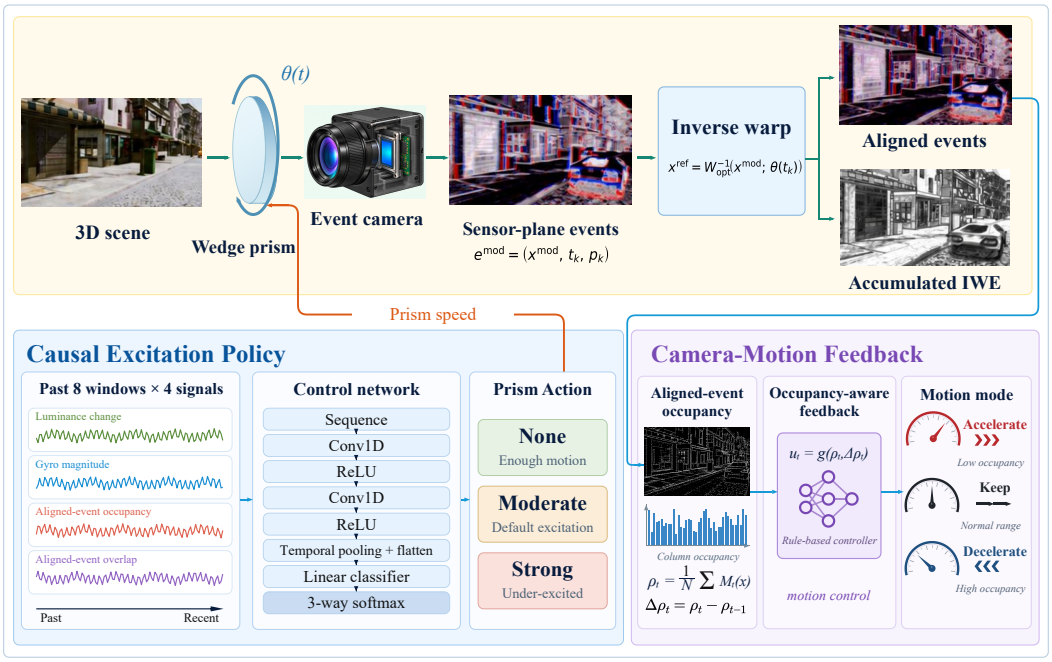


Figure 2: **Overview of EIRM.** EIRM combines learned optical excitation with rule-based camera-motion feedback in a causal sensing loop. A temporal policy selects the prism excitation from recent visual, inertial, and aligned-event observations, while aligned-event occupancy and its temporal change regulate camera motion. Timestamp-aware inverse warping aligns the induced events.

3 METHOD

EIRM forms a causal active-sensing loop that combines a learned temporal policy for optical excitation with rule-based camera-motion feedback, as illustrated in Fig. 2. At each control interval, the policy uses a fixed-length history of completed sensing windows to select the prism action for the next interval, while aligned-event feedback regulates the progression of the controlled camera trajectory (Sec. 3.1). The resulting events are then aligned through timestamp-aware compensation (Sec. 3.2). To evaluate whether this adaptive sensing produces useful rather than merely abundant events, we further introduce Target Event Observability (TEO; Sec. 3.3) and a hybrid sim-and-real benchmark spanning diverse camera–object relative-motion regimes (Sec. 3.4).

3.1 CAUSAL OPTICAL EXCITATION AND MOTION REGULATION

We divide a sequence into consecutive control windows $\mathcal{I}_n = (\tau_{n-1}, \tau_n]$ and summarize each completed window by $\mathbf{s}_n = [d_n^L, d_n^{\text{imu}}, d_n^{\text{occ}}, d_n^{\text{ov}}]^\top$. Here, d_n^L measures log-luminance variation, d_n^{imu} normalized gyroscope magnitude, d_n^{occ} the spatial occupancy of aligned events, and d_n^{ov} their support overlap with the preceding window. Specifically, for aligned-event support $\mathcal{S}_n^{\text{ref}}$ and valid image region Ω_v, $d_n^{\text{occ}} = |\mathcal{S}_n^{\text{ref}}|/|\Omega_v|$, while d_n^{ov} is the intersection-over-union between $\mathcal{S}_n^{\text{ref}}$ and $\mathcal{S}_{n-1}^{\text{ref}}$.

Adaptive Prism Excitation. A causal temporal policy maps the eight most recent completed sensing states to the prism action for the next interval. The action space is$\mathcal{A}_p = \{0, 180, 360\}$ RPM, corresponding to no, moderate, and strong optical excitation, with 0 RPM denoting passive event sensing without prism-induced motion.The policy is trained by imitation from offline privileged supervision: for each training window, the three candidate actions are evaluated using image-wide edge F1, and the best action serves as the classification target.Privileged measurements are used only for label construction; inference relies solely on completed sensing states and remains causal.Architecture, teacher construction, and optimization details can be seen in the Appendix.

Event-driven camera-motion regulation. A complementary rule-based controller regulates progression along the controlled camera trajectory using aligned-event occupancy and its temporal change.For the next interval, it selects one of three motion modes: *accelerate*, *keep*, or *decelerate*.Low and non-increasing occupancy permits acceleration, while high or rapidly increasing occupancy triggers deceleration; otherwise, the current rate is maintained.The prism policy and motion controller thus regulate complementary sources of sensing motion: optical excitation and controlled camera motion.The selected prism action defines the angular trajectory $\theta(t)$ used for active image formation and timestamp-aware compensation in Sec. 3.2. Controller thresholds, rate levels, and dwell constraints can be seen in the Appendix.

3.2 Pre-Event Optical Modulation and Timestamp-Aware Compensation

Optical modulation is applied *before* event generation. Let $Y(\mathbf{x}, t)$ denote a temporally dense luminance trajectory and $W_{\mathrm{opt}}(\mathbf{x}; \theta)$ the calibrated image-plane transformation induced by the wedge prism. For prism orientation $\theta(t)$, the event sensor directly receives

$$Y_{\mathrm{opt}}(\mathbf{x}, t) = Y\left(W_{\mathrm{opt}}^{-1}(\mathbf{x}; \theta(t)), t\right), \tag{1}$$

from which events are generated. Because prism motion changes the temporal intensity trajectory observed at each individual pixel, it can alter whether an event is triggered as well as its timestamp, location, and polarity; these effects cannot be reproduced by simple post-hoc warping.

For an event $e_k^{\mathrm{mod}} = (\mathbf{x}_k^{\mathrm{mod}}, t_k, p_k)$, the known prism trajectory allows the prism-induced image displacement to be explicitly removed at the exact event timestamp:

$$\mathbf{x}_k^{\mathrm{ref}} = W_{\mathrm{opt}}^{-1}\left(\mathbf{x}_k^{\mathrm{mod}}; \theta(t_k)\right), \qquad e_k^{\mathrm{ref}} = (\mathbf{x}_k^{\mathrm{ref}}, t_k, p_k). \tag{2}$$

Timestamp and polarity remain unchanged. Using $\theta(t_k)$ rather than a single interval-level prism pose avoids residual displacement between events generated at different times and yields a temporally consistent common reference-plane stream for feedback and evaluation.Temporal densification, optical calibration, and event-generation details can be seen in the Appendix.

3.3 Target Event Observability

Event count alone does not characterize sensing quality: sparse streams may miss important structure, while dense streams may contain substantial redundancy. We therefore introduce *Target Event Observability* (TEO) to measure useful target observation beyond raw event abundance.

Let $\mathcal{T}$ denote the target region, $\mathcal{G} \subseteq \mathcal{T}$ its edge set, and $\mathcal{M}$ the support of aligned events. We characterize spatial observability using target-edge coverage C and event purity P:

$$C = \frac{|\mathcal{G} \cap \mathcal{M}|}{|\mathcal{G}|}, \qquad P = \frac{|\mathcal{G} \cap \mathcal{M}|}{|\mathcal{M} \cap \mathcal{T}|}. \tag{3}$$

C measures the fraction of target-edge structure that is observed, whereas P measures how strongly in-target events concentrate on informative boundaries; background events therefore do not enter its denominator. To capture temporal persistence, let $\tau \in [0, 1]$ denote the fraction of temporal slices containing target-edge observations, and define

$$\mathrm{TEO} = \frac{2CP}{C + P}\,\tau, \tag{4}$$

with the corresponding score set to zero whenever the required support is empty or $C + P = 0$. TEO is high only when informative target structure is broadly covered, event responses remain concentrated on relevant boundaries, and these observations persist consistently throughout the evaluated temporal window rather than appearing only transiently.

3.4 Hybrid Relative-Motion Benchmark

To evaluate the motion-adaptive behavior introduced in Sec. 3.1, we construct a hybrid sim-and-real benchmark with controlled camera–object relative-motion regimes. In Blender, we independently

control camera and object motion under four regimes: *camera-only* and *object-only* isolate the contribution of each motion source, *co-motion* synchronizes camera and object motion to suppress relative image displacement, and *counter-motion* drives them oppositely to amplify it. These regimes probe whether additional excitation is beneficial when natural relative motion is weak and unnecessary when it is already strong. Blender further provides target masks, and TEO (Sec. 3.3) is therefore evaluated on the Blender subset.

We additionally include reconstructed DL3DV scenes (Ling et al., 2024), which are built from real captures and re-rendered under controlled camera trajectories, as well as recorded DSEC (Gehrig et al., 2021) sequences, a real-world driving dataset with synchronized RGB and IMU streams that provides natural visual–inertial dynamics. Because EIRM modulates image formation before event generation (Sec. 3.2), prerecorded native events cannot reflect the imposed optical excitation. We therefore regenerate events from temporally dense visual signals using the same pre-event modulation pipeline across all sources. The inertial input is derived from the known camera motion for Blender and DL3DV, and taken from the recorded IMU stream for DSEC. Scene construction, trajectory generation, inertial-signal derivation, and target-mask details can be seen in the Appendix.

4 Experiments

We evaluate EIRM on the hybrid benchmark of Sec. 3.4, progressing from sensing behavior to mechanism analysis. We first study how optical excitation interacts with different camera–object relative motions, then evaluate whether the resulting events provide useful target observations through TEO, and finally analyze the contributions of adaptive excitation and timestamp-aware compensation. Sec. 4.1 details the training and evaluation protocol.

4.1 Experimental Setup

Training and Evaluation. We evaluate on the hybrid sim-and-real benchmark introduced in Sec. 3.4. Controlled Blender scenes provide independently configurable camera and object motion together with target masks, covering the four relative-motion regimes: *camera-only*, *object-only*, *co-motion*, and *counter-motion*. Reconstructed DL3DV scenes (Ling et al., 2024) provide realistic appearance under controlled rendering and camera trajectories, while recorded DSEC RGB–IMU trajectories are evaluated in the Appendix. The optical-excitation policy is trained exclusively on training sequences and selected only on a separate validation split; all reported quantitative results are evaluated on held-out sequences excluded from teacher construction, feature normalization, and checkpoint selection. Split and training details can be seen in the Appendix.

Baselines and Metrics. DVS-Voltmeter (Lin et al., 2022) and v2e (Hu et al., 2021) are passive event simulators applied to matched source RGB sequences and timestamps, and do not share our event-generation backend. Fixed optical policies hold the prism speed constant throughout each sequence; Fixed-X denotes a constant prism speed of X RPM and, together with EIRM, uses the same timestamped prism-warp and threshold-event simulation pipeline, with optical calibration, event generation, temporal sampling, compensation, and evaluation held fixed. The controlled motion–excitation sweep considers $0/180/360$ RPM, corresponding to passive, moderate, and strong optical excitation, respectively. The 0 RPM setting introduces no prism-induced motion and represents the same passive sensing regime as Fixed-0; it is therefore covered by the passive simulator baselines. Fixed-180 serves as the nominal active-sensing baseline, and Fixed-360 is additionally included in the policy and accumulation analyses. We report TEO following Eqs. (3) to (4), together with image-wide edge F1, precision, recall, event activity, and mean prism speed, computed per temporal window and averaged over the corresponding evaluation split or motion regime. Aligned-accumulation fidelity is measured using FSIM (Zhang et al., 2011) and LPIPS (Zhang et al., 2018) against an RGB-derived normalized log-luminance gradient reference. Throughout, the best and second-best results in each column are shown in bold and italics, respectively. Metric definitions and evaluation settings can be seen in the Appendix.

Implementation. All fixed and adaptive optical-sensing results are produced by a self-developed, CUDA-accelerated prism-warp and threshold-event simulation pipeline, with event simulation, compensation, and evaluation running on an NVIDIA RTX 4090 GPU. Further implementation details can be seen in the Appendix.

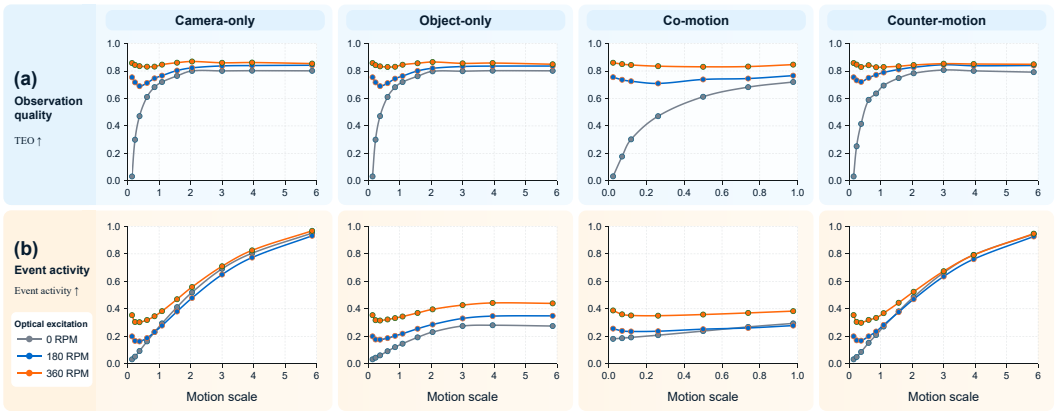


Figure 3: **Motion-dependent optical excitation.** TEO (a) and normalized event activity (b) under fixed prism speeds across the four relative-motion regimes. Increasing excitation can raise activity even after useful target observation has saturated.

Table 1: **Same-source downstream comparison on the held-out controlled benchmark.** Mean F1 averages the four relative-motion regimes. All metrics are higher-is-better.

	F1						
Method	Mean	Camera-only	Object-only	Co-motion	Counter-motion	Precision	Recall
DVS-Voltmeter	0.3291	0.3407	0.6545	0.0000	0.3213	**0.3746**	0.3923
v2e	0.3442	0.3350	**0.7230**	0.0000	0.3186	*0.3593*	0.4610
Fixed-180	*0.3907*	*0.3750*	0.4286	*0.3575*	**0.4016**	0.2924	*0.6683*
EIRM(Ours)	**0.3930**	**0.3759**	0.4286	**0.3752**	*0.3921*	0.2963	**0.6770**

4.2 Motion-Dependent Active Event Sensing

Motion–Excitation Response. We first isolate optical excitation from the learned controller by replaying matched camera–object trajectories with fixed prism speeds while keeping the active event-generation and evaluation pipeline unchanged. Fig. 3 reports TEO together with normalized event activity. Each point averages results from three seed-indexed rendered scene realizations.When target-relative image motion is weak, added optical motion exposes structure that would otherwise be poorly observed. As natural relative motion increases, the gain in TEO saturates while event activity can continue to grow without improving target coverage proportionally. The preferred excitation therefore depends on the camera–object motion regime rather than on event count alone.

Same-Source Structural Comparison. Table 1 evaluates image-wide edge reconstruction on matched source sequences, timestamps, temporal windows, and scoring procedures. The regimes expose complementary behavior. Under co-motion, the passive simulators retain nonzero event activity but the evaluated windows contain no target-mask events; edge overlap is therefore zero, while optical excitation restores measurable structure. Under object-only motion, by contrast, target motion already supplies strong temporal contrast and the passive simulators obtain higher image-wide F1. This difference motivates separating global edge reconstruction from target-level observability.

Qualitative Event Observability. Fig. 4 compares matched Blender and reconstructed DL3DV examples. When relative motion is weak or degenerate, passive simulation can preserve background activity while informative target or scene structure remains poorly represented in the resulting event stream. Optical excitation introduces controlled image-plane motion before event formation, and timestamp-aware compensation keeps these responses spatially organized across the image.

4.3 Target Observability Across Motion Regimes

Image-wide reconstruction does not indicate whether the target itself is sufficiently observed. This distinction is especially important under co-motion, where background structure can continue gen-

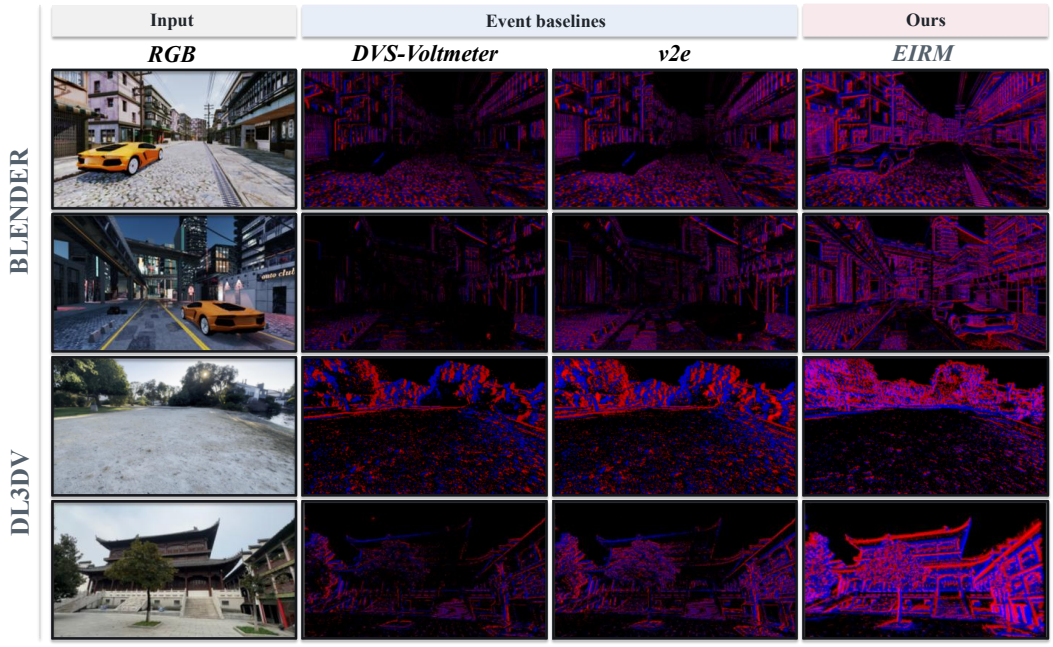


Figure 4: **Qualitative event observability.** Matched examples from controlled Blender scenes and reconstructed DL3DV scenes across representative motion conditions. Active excitation is most useful when the original relative motion cannot reliably reveal coherent scene structure.

erating events while the target has little relative image motion and therefore remains weakly represented despite nonzero overall event activity. Because the masks and reference edges are defined independently by the benchmark rather than by the event generator, TEO is evaluated with the same protocol across all methods, enabling a direct comparison of target-level observability.

Table 2: **TEO across relative-motion regimes on the Blender subset.** Higher values indicate better target observability; zero under co-motion indicates windows with no target-mask events.

Method	Camera-only	Object-only	Co-motion	Counter-motion	Average
DVS-Voltmeter	0.6633	0.6597	0.0000	0.6334	0.4891
v2e	0.7209	*0.7156*	0.0000	0.6945	0.5328
Fixed-180	*0.7669*	**0.7629**	*0.7724*	**0.7889**	*0.7728*
EIRM(Ours)	**0.7713**	**0.7629**	**0.7981**	*0.7765*	**0.7772**

Table 2 reports TEO across the four relative-motion regimes. Co-motion most clearly separates event abundance from useful target observation. Passive DVS-Voltmeter and v2e emit background events, but their stored co-motion windows contain no target-mask events, causing $C = P = \tau = 0$ and therefore zero TEO. Optical excitation restores target-relative temporal contrast by introducing motion absent from the natural camera–object interaction. In the remaining regimes, camera or object motion provides informative temporal variation, reducing the need for additional excitation. This pattern indicates that stronger excitation is most useful when natural relative motion is insufficient, whereas its marginal benefit decreases once useful motion is present. EIRM adapts to these sensing conditions while maintaining strong target observability without uniformly strong excitation.

4.4 Component Ablations and Representation Fidelity

Adaptive Optical Excitation. Table 3(a) compares fixed and adaptive optical-excitation policies under the same camera–object relative-motion conditions. Fixed policies maintain a constant prism speed, whereas EIRM selects the next excitation level from the completed sensing history, allowing

optical modulation to adapt as relative motion changes. EIRM achieves the strongest F1 while operating near the moderate-excitation regime rather than defaulting to the strongest optical action. This is consistent with Fig. 3: stronger modulation is not uniformly beneficial across motion conditions.

Timestamp-Aware Compensation. To isolate inverse-warp timing, we compare no compensation, one prism state per interval, and per-event compensation. These variants operate on the identical event stream produced by EIRM on reconstructed DL3DV scenes; only the temporal precision of the inverse warp changes. In Table 3(b), *Cov.* denotes whole-image RGB-edge coverage and differs from target-level coverage C in TEO. All three metrics improve monotonically from no compensation to interval-level and per-event compensation, indicating that inverting the prism state at each event timestamp removes residual displacement left unresolved by a single interval-level pose.

Table 3: **Optical-excitation and compensation ablations. Left:** fixed and adaptive optical-excitation policies on the held-out controlled benchmark. **Right:** inverse-warp timing on reconstructed DL3DV scenes. All metrics are higher-is-better.

(a) Optical excitation policy

Method	F1	Mean RPM
Fixed-180	*0.3907*	180.0
Fixed-360	0.3839	360.0
EIRM(Ours)	**0.3930**	197.1

(b) Timestamp-aware compensation

Compensation	F1	Prec.	Cov.
None	0.5195	0.3945	0.7924
Interval	*0.5696*	*0.4421*	*0.8332*
Per-event	**0.5890**	**0.4572**	**0.8603**

Accumulation Fidelity. We finally compare timestamp-aligned event accumulations with an RGB-derived normalized log-luminance gradient reference using FSIM and LPIPS (Table 4). FSIM measures structural similarity in local feature organization, whereas LPIPS captures perceptual discrepancy between the accumulated event structure and the RGB-derived reference. Together, they provide a complementary measure of structural fidelity.

Table 4: **Aligned event-accumulation fidelity.** FSIM and LPIPS are computed against the RGB-derived structural reference on the full split and the co-motion subset.

Method	FSIM ↑	LPIPS ↓	Co-motion FSIM ↑	Co-motion LPIPS ↓
DVS-Voltmeter (Lin et al., 2022)	0.4833	0.7916	0.5464	0.6497
v2e (Hu et al., 2021)	0.4935	0.7701	0.5506	0.6132
Fixed-180	**0.6635**	**0.4595**	**0.6371**	*0.4654*
Fixed-360	0.6254	0.4681	0.6185	**0.4582**
EIRM(Ours)	*0.6555*	*0.4649*	*0.6268*	0.4679

All optically excited variants remain substantially closer to the structural reference than the passive baselines, showing that the added activity is not merely unstructured event density. Fixed-180 gives the strongest full-split fidelity, while Fixed-360 achieves the lowest co-motion LPIPS; EIRM remains close to these fixed settings rather than consistently dominating every fidelity metric. Together, the experiments show that excitation is most useful when natural relative motion is insufficient, while timestamp-aware compensation preserves the geometry of the induced events.

5 Conclusion

We presented EIRM, an event-sensing framework for actively controlled acquisition that regulates optical excitation and controlled camera motion before event generation, guided by inertial measurements, and compensates the induced optical motion at each event's timestamp through timestamp-aware inverse warping. Experiments on a hybrid sim-and-real benchmark show that event quantity alone does not imply target observability: active sensing recovers target observation when natural motion is insufficient, while stronger optical motion is not uniformly beneficial across conditions.

AI USE STATEMENT

Generative AI tools were used to assist with software implementation and debugging, manuscript drafting and language polishing, and the layout and visual refinement of scientific figures. AI tools were not used to generate experimental results, formulate scientific claims, design the methodology or experiments, or interpret the reported results. All AI-assisted code, text, and figures were manually reviewed, revised, and verified by the authors. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details necessary to reproduce our results in the main paper and appendix. The optical excitation mechanism, motion-feedback controller, and timestamp-aware compensation are described in the method section, while implementation details, simulator and event-generation configurations, and evaluation protocols are provided in the appendix. The relevant code and configuration files will be released upon publication.

REFERENCES

Dora E. Angelaki and Bernhard J. M. Hess. Self-motion-induced eye movements: Effects on visual acuity and navigation. *Nature Reviews Neuroscience*, 6(12):966–976, 2005. doi: 10.1038/nrn1804.

Tobi Delbruck, Rui Graca, and Marcin Paluch. Feedback control of event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1324–1332, 2021.

Chaoran Feng, Wangbo Yu, Xinhua Cheng, Zhenyu Tang, Junwu Zhang, Li Yuan, and Yonghong Tian. Ae-nerf: Augmenting event-based neural radiance fields for non-ideal conditions and larger scenes. volume 39, pp. 2924–2932, 2025.

Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3867–3876, 2018. doi: 10.1109/CVPR.2018.00407.

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, 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, 2021. doi: 10.1109/LRA.2021.3068942.

Botao He, Ze Wang, Yuan Zhou, Jingxi Chen, Chahat Deep Singh, Haojia Li, Yuman Gao, Shaojie Shen, Kaiwei Wang, Yanjun Cao, Chao Xu, Yiannis Aloimonos, Fei Gao, and Cornelia Fermüller. Microsaccade-inspired event camera for robotics. *Science Robotics*, 9(90):eadj8124, 2024. doi: 10.1126/scirobotics.adj8124.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic DVS events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1312–1321, 2021. doi: 10.1109/CVPRW53098.2021.00144.

Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Computer Vision – ECCV 2022*, volume 13674 of *Lecture Notes in Computer Science*, pp. 624–642. Springer, 2022. doi: 10.1007/978-3-031-19781-9_36.

Xutao Kuang, Martina Poletti, Jonathan D. Victor, and Michele Rucci. Temporal encoding of spatial information during active visual fixation. *Current Biology*, 22(6):510–514, 2012. doi: 10.1016/j.cub.2012.01.050.

Yuchen Li, Chaoran Feng, Zhenyu Tang, Kaiyuan Deng, Wangbo Yu, Yonghong Tian, and Li Yuan. Gs2e: Gaussian splatting is an effective data generator for event stream generation, 2025. URL https://arxiv.org/abs/2505.15287.

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, 2008. doi: 10.1109/JSSC.2007.914337.

Songnan Lin, Ye Ma, Zhenhua Guo, and Bihan Wen. DVS-Voltmeter: Stochastic process-based event simulator for dynamic vision sensors. In *Computer Vision – ECCV 2022*, volume 13667 of *Lecture Notes in Computer Science*, pp. 578–593. Springer, 2022. doi: 10.1007/978-3-031-20071-7_34.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Weng Fei Low and Gim Hee Lee. Robust e-nerf: Nerf from sparse & noisy events under non-uniform motion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18289–18300. IEEE, 2023.

Wenzong Ma, Zhuoxiao Li, Jinjing Zhu, Tongyan Hua, Kanghao Chen, Zidong Cao, Da Yang, Peilun Shi, Yibo Zhou, Wufan Zhao, et al. Skyevents: A large-scale event-enhanced uav dataset for robust 3d scene reconstruction. In *The Fourteenth International Conference on Learning Representations*, 2026.

Abhishek Mishra, Rohan Ghosh, Jose C. Principe, Nitish V. Thakor, and Sunil L. Kukreja. A saccade based framework for real-time motion segmentation using event based vision sensors. *Frontiers in Neuroscience*, 11:83, 2017. doi: 10.3389/fnins.2017.00083.

Elias Mueggler, Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. Continuous-time visual-inertial odometry for event cameras. *IEEE Transactions on Robotics*, 34(6):1425–1440, 2018. doi: 10.1109/TRO.2018.2858287.

Garrick Orchard, Ajinkya Jayawant, Gregory K. Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in Neuroscience*, 9:437, 2015. doi: 10.3389/fnins.2015.00437.

Roberto Pellerito, Nico Messikommer, Giovanni Cioffi, Marco Cannici, and Davide Scaramuzza. Motion-aware event suppression for event cameras. In *Robotics: Science and Systems*, 2026.

Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: An open event camera simulator. In *Proceedings of the 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pp. 969–982. PMLR, 2018.

Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate SLAM? combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios. *IEEE Robotics and Automation Letters*, 3(2):994–1001, 2018. doi: 10.1109/LRA.2018.2793357.

Michele Rucci, Ramon Iovin, Martina Poletti, and Fabrizio Santini. Miniature eye movements enhance fine spatial detail. *Nature*, 447(7146):852–855, 2007. doi: 10.1038/nature05866.

Simone Testa, Giacomo Indiveri, and Silvio P. Sabatini. A bio-inspired neuromorphic active vision system based on fixational eye movements. In *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2020. doi: 10.1109/ISCAS45731.2020.9180439.

Simone Testa, Silvio P. Sabatini, and Andrea Canessa. Active fixation as an efficient coding strategy for neuromorphic vision. *Scientific Reports*, 13:7445, 2023. doi: 10.1038/s41598-023-34508-x.

Xiaoshan Wu, Yifei Yu, Xiaoyang Lyu, Yihua Huang, Bo Wang, Baoheng Zhang, Zhongrui Wang, and Xiaojuan Qi. Eag3r: Event-augmented 3d geometry estimation for dynamic and extreme-lighting scenes. *Advances in Neural Information Processing Systems*, 38:73314–73342, 2026.

Amirreza Yousefzadeh, Garrick Orchard, Teresa Serrano-Gotarredona, and Bernabe Linares-Barranco. Active perception with dynamic vision sensors: Minimum saccades with optimum recognition. *IEEE Transactions on Biomedical Circuits and Systems*, 12(4):927–939, 2018. doi: 10.1109/TBCAS.2018.2834428.

Wangbo Yu, Chaoran Feng, Jianing Li, Aofan Zhang, Zhenyu Tang, Mingyi Guo, Wei Zhang, Zhengyu Ma, Li Yuan, and Yonghong Tian. Ea3d: Event-augmented 3d diffusion for generalizable novel view synthesis. In *ICLR 2026*, 2026.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. doi: 10.1109/TIP.2011.2109730.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Alex Zihao Zhu, Nikolay Atanasov, and Kostas Daniilidis. Event-based visual inertial odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5391–5399, 2017.

Alex Zihao Zhu, Dinesh Thakur, Tolga Ozaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multi vehicle stereo event camera dataset: An event camera dataset for 3d perception. *IEEE Robotics and Automation Letters*, 3(3):2032–2039, 2018. doi: 10.1109/LRA.2018.2800793.

A Modeling Assumptions and Optical Details

A.1 Modeling Assumptions and Scope

Calibration and synchronization. We assume known camera intrinsics, a calibrated rotation from the IMU frame to the camera frame, and a calibrated effective optical deflection angle. Visual, inertial, and optical-state signals are represented on a common timeline. Within the valid image support Ω_v, the calibrated optical mapping is assumed to remain locally invertible over all admissible actuator states.

Observable motion. The inertial cue used by EIRM describes the rotational component of ego-motion observable from the gyroscope. Translation-induced image motion depends on scene depth and is therefore not inferred from the gyroscope alone. Independently moving objects and illumination changes remain part of the observed visual signal rather than quantities explicitly canceled by the controller.

Actuator model. The rotating optical element provides a single controllable degree of freedom, with bounded angular velocity and acceleration. EIRM therefore regulates only the image-plane motion component attainable by the calibrated optical element and does not assume arbitrary two-dimensional retinal stabilization. In the simulation experiments, the commanded optical trajectory is treated as the realized trajectory and is available continuously for event formation and inverse alignment.

Event simulation scope. The active-sensing experiments use the timestamped prism-warp and threshold-event simulation pipeline. DVS-Voltmeter and v2e retain their native passive event-simulation implementations. Comparisons against these methods are therefore matched at the source RGB sequence, timestamps, temporal windows, and evaluation procedure, while fixed and adaptive active policies additionally share the same optical mapping, event-generation backend, and compensation pipeline.

A.2 Optical Mapping and Calibration

Let $\mathbf{K}$ denote the camera intrinsic matrix and $\overline{\mathbf{x}}$ the homogeneous coordinate associated with image location $\mathbf{x}$. The corresponding unit camera ray is

$$\mathbf{r}(\mathbf{x}) = \frac{\mathbf{K}^{-1}\overline{\mathbf{x}}}{\|\mathbf{K}^{-1}\overline{\mathbf{x}}\|_2}. \tag{5}$$

For a calibrated effective deflection angle δ, the rotating optical element induces the paraxial deflection

$$\mathbf{q}(\theta) = \delta \begin{bmatrix} \cos\theta & \sin\theta & 0 \end{bmatrix}^\top. \tag{6}$$

Using perspective projection $\pi(\cdot)$, the induced image-plane mapping is

$$W_{\text{opt}}(\mathbf{x};\theta) = \pi\left(\mathbf{K}\frac{\mathbf{r}(\mathbf{x}) + \mathbf{q}(\theta)}{\|\mathbf{r}(\mathbf{x}) + \mathbf{q}(\theta)\|_2}\right). \tag{7}$$

The effective angle δ absorbs the small-angle refractive parameters of the optical element and is estimated during geometric calibration. The same calibrated mapping is used for pre-event optical modulation and timestamp-aware inverse alignment.

B Benchmark Construction

This section details the construction of the hybrid benchmark introduced in Sec. 3.4: the Blender scenes and their relative-motion trajectories, the inertial-signal derivation, target masks and reference edges, the DL3DV reconstruction and re-rendering, the DSEC sequence selection, and the shared event-regeneration configuration.

B.1 Blender Scenes and Trajectories

Rendering. All Blender scenes are rendered in Blender 5.2.1 LTS with the EEVEE engine at 960×540 resolution, 480 RGB frames at 60 Hz (8 s per clip), lossless PNG, without motion blur. The camera uses a 16 mm lens on a 36 mm horizontal sensor, giving $f_x = f_y = 426.67$ px, $(c_x, c_y) = (480, 270)$, a pinhole model without distortion. Each clip is accompanied by an inertial stream at 1000 Hz (Sec. B.2).

Scene instances. The controlled split uses three scene instances: Scene_002 *Residential Avenue* (train), Scene_004 *Neon Auto District* (validation), and Scene_005 *Cyberpunk Boulevard* (test). Each instance is built from a different source city asset, and the test scene originates from a different source blend than the training and validation scenes, so the splits are source-disjoint. Both the ego and the neighbor vehicle are instances of the same glTF car asset; the camera is mounted 1.20 m ahead and 1.25 m above the ego. The seed index of each realization is the scene identifier itself (2, 4, 5): the three realizations are therefore distinct scene instances rather than stochastic reruns of one scene.

Trajectories. Each scene follows a scripted longitudinal speed profile $p(u)$, $u \in [0, 1]$, with nominal speed $\frac{\text{path length}}{8\,\text{s}}\,p(u)$: WAVE_ONE ($a$=0.45, 2.06–5.44 m/s over a 30 m path), TWO_PULSE (a=0.30, 3.5–6.5 m/s, 40 m), and BRAKE_ACCEL (a=0.45, 1.93–5.08 m/s, 28 m). The four relative-motion regimes are realized by freezing or superposing the camera/ego and neighbor trajectories:

- *camera-only*: the neighbor vehicle is frozen at $u = 0$;
- *object-only*: the camera and ego are frozen at $u = 0$, while the neighbor travels its path with an additional ± 2.5 m lateral sinusoid;

- *co-motion*: both move with locked relative pose (residual longitudinal offset below 4×10^{-6} m), suppressing relative image displacement;
- *counter-motion*: the camera follows the co-motion path while the neighbor adds the same ± 2.5 m sinusoid on top of the locked pose, amplifying relative displacement.

In all cases the maximum longitudinal relative displacement is bounded by the 2.5 m sinusoid amplitude. Three scenes $\times$ four regimes give the 12 clips of the controlled split (Sec. E).

B.2 Inertial-Signal Derivation

For Blender, the inertial stream is synthesized alongside the known trajectory at 1000 Hz. The longitudinal acceleration component follows the second derivative of the speed profile and is present only when the camera moves (i.e., in all regimes except object-only); the gyroscope channels are small-amplitude deterministic sinusoids with scene-specific phases. No additional noise or filtering is applied. The window-level inertial feature is the mean gyroscope ℓ_2 norm over each RGB frame interval, with linear interpolation for empty intervals; features are min–max normalized using training-scene statistics and clipped. For DL3DV, the inertial signal is obtained by pose differencing on the reconstructed camera trajectory and is therefore pose-derived rather than sensor-recorded. For DSEC, the recorded IMU stream is used directly.

B.3 Target Masks and Reference Edges

Target masks are generated by projecting the 3D axis-aligned bounding box of the neighbor vehicle ($2.0 \times 4.6 \times 1.15$ m) onto the image plane and filling the resulting rectangle with 255 (8-bit); pixels above 127 are foreground. A window is scored only when the projected mask covers at least 5% of the image. Reference edges are computed from the reference image as the gradient magnitude of log-clipped intensity, thresholded at the 75th percentile of nonzero magnitudes; target edges are the subset falling inside the target mask.

B.4 DL3DV Reconstruction and Re-rendering

We use five DL3DV scenes (Ling et al., 2024). The reconstructed scenes are re-rendered under controlled camera trajectories following the standard GS2E pipeline (Li et al., 2025). The rendered RGB sequences contain 40 frames per clip at 960×540 resolution, with timestamps following the nominal 30 fps capture interval.

Since event generation requires temporally dense visual signals, we further densify the rendered RGB sequences using Practical-RIFE (Huang et al., 2022). Specifically, each 40-frame sequence is interpolated to 480 frames before event regeneration. The resulting dense RGB sequence is used as input to the same prism-warp, threshold-event generation, and timestamp-aware compensation pipeline as other evaluated sources. RIFE inference is performed with GPU acceleration for efficient preprocessing.The compensation ablation of Sec. 4.4 evaluates at width 160; the effective prism deflection for DL3DV is $2.0°$.

B.5 DSEC Sequence Selection

Recorded evaluation uses the DSEC *thun_00_a* sequence (Gehrig et al., 2021): 120 frames at 30 fps (4 s), 50 ms event-generation windows, and 640×480 simulation resolution. Image timestamps are taken from the recorded timestamp file; the IMU stream is extracted for the same interval. The native DSEC events are retained only as an unaligned recorded reference (Sec. H).

B.6 Event-Regeneration Configuration

Because optical modulation acts before event formation, all compared sensing variants on all three sources regenerate events from the temporally dense visual signal with the same calibrated prism-warp and threshold-event pipeline: contrast threshold 0.2, no refractory period, prism phase subdivided so that no sub-step exceeds the configured maximum phase step, and event timestamps obtained by linear interpolation of the log-intensity threshold crossing.For DL3DV, temporal densification is performed before prism modulation usingPractical-RIFE interpolation from 40 rendered

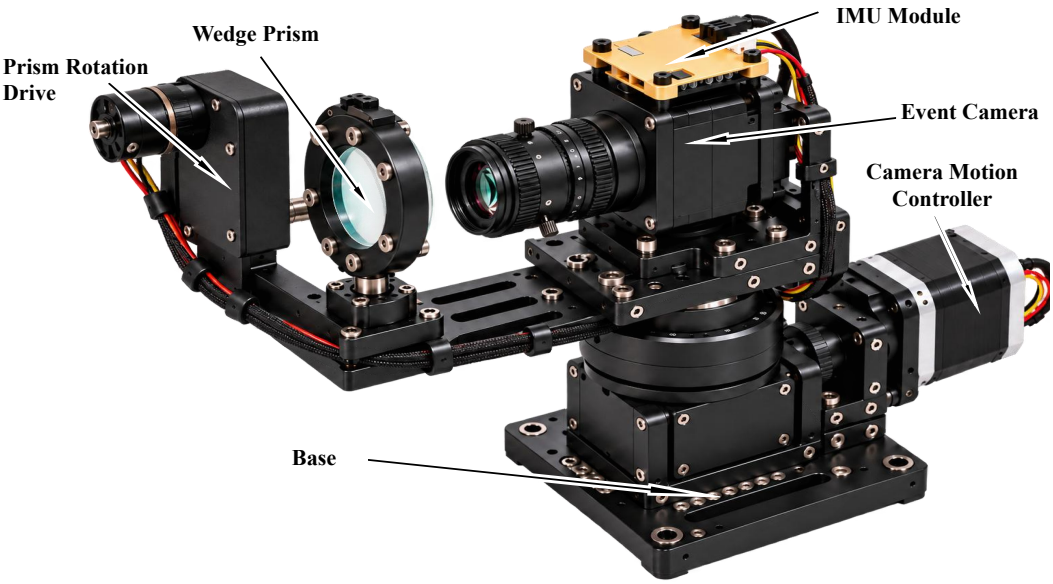


Figure 5: **Rotating-wedge-prism hardware prototype.** The prism rotation drive and the camera motion controller realize the two sensing-motion branches of EIRM; the IMU module is rigidly attached to the event camera.

Table 5: **Prototype components.** The host executes the causal sensing policy; the microcontroller closes the low-level control loop and streams inertial and motor state to the host.

Component	Model	Role
Host	Intel NUC 13	runs the optical-excitation policy, camera-motion feedback, and timestamp-aware compensation
Microcontroller	Damiao H7	executes prism-speed commands; samples the IMU and streams inertial and motor state to the host
Motor	DJI RoboMaster M2006	drives the rotating wedge prism
Event camera	SilkyEvCam VGA	asynchronous sensing front-end; Prophesee PPS3MVCD (Gen3.1), 640 × 480, microsecond timestamps, USB 3.0
IMU	Bosch BMI088 breakout board	6-axis inertial sensing; rigidly attached to the camera body
Optical element	wedge prism (single plate)	rotated by the motor; introduces the calibrated image-plane deflection

frames to 480 frames. The effective deflection is 0.5° for Blender and 2.0° for DL3DV. Passive baselines keep their native implementations instead: v2e (positive/negative threshold 0.2, noise threshold 0.03, leak rate 0.1 Hz) and DVS-Voltmeter (DVS346 model, CPU). Comparisons are therefore matched at the source sequence, timestamps, temporal windows, and scoring procedure (Sec. E).

C Hardware Prototype

Fig. 5 shows the assembled prototype. A single wedge prism is held in a rotary mount driven by a motor, forming the optical-excitation branch of Sec. D.2: the prism-speed commands selected by the policy are applied to this drive. The camera sits on a motorized base that realizes the camera-motion feedback branch of Sec. D.4. An IMU module is bolted directly onto the camera body so that inertial and visual signals share one rigid frame, as assumed in Sec. A. The control chain runs across two compute levels: a host machine executes the causal sensing policy, while a microcontroller closes the low-level motor loop and streams inertial and motor state upward. Table 5 summarizes the components.

The sensing front-end is the SilkyEvCam VGA (CenturyArks), an event camera built around the Prophesee PPS3MVCD (Gen3.1) sensor with 640×480 resolution and microsecond event timestamping; its native resolution coincides with the 640×480 simulation resolution used for the recorded DSEC evaluation (Sec. H).

The effective deflection angle δ of Sec. A.2 is estimated by imaging a calibration target at known prism orientations and fitting the paraxial model of Eq. (7) to the measured displacements. Visual, inertial, and optical-state signals are time-stamped on a common clock before entering the control pipeline. The measured optical calibration and synchronized sensing configuration are used throughout the evaluated EIRM sensing pipeline.

D Policy Learning and Control Implementation

D.1 Causal Sensing State

The optical-excitation policy operates only on completed sensing windows. For window n, the causal state is

$$\mathbf{s}_n = \left[d_n^L, d_n^{\mathrm{imu}}, d_n^{\mathrm{occ}}, d_n^{\mathrm{ov}}\right]^\top, \tag{8}$$

where d_n^L is the 0.9 quantile of $|\Delta \log I|/\Delta t$ over the completed window, d_n^{imu} is the mean gyroscope ℓ_2 magnitude, d_n^{occ} is the unique-pixel occupancy of the aligned event support, and d_n^{ov} is the Jaccard overlap between the aligned supports of windows n and $n-1$. The IMU magnitude is min–max normalized using training data, after which all four features are standardized using training-set statistics and clipped to $[-5, 5]$.

The policy consumes the previous eight completed states,

$$\mathbf{S}_n = [\mathbf{s}_{n-7}, \ldots, \mathbf{s}_n], \tag{9}$$

and predicts the optical action for the following interval. No future visual, inertial, or event information is used.

D.2 Optical-Excitation Policy

The causal policy uses two one-dimensional temporal convolution blocks. The first maps 4 input channels to 32 channels with kernel size 3 and dilation 1; the second maps 32 channels to 32 channels with kernel size 3 and dilation 2. Both use left causal padding, GroupNorm with one group, and ReLU. The final temporal feature is passed through a $32 \to 16$ linear layer with ReLU and a $16 \to 3$ classifier. No dropout is used. The three logits correspond to 360, 180, and 0 RPM, respectively.

The general optical action space is

$$\mathcal{A}_p = \{0, 180, 360\}\ \mathrm{RPM}, \tag{10}$$

representing no, moderate, and strong optical excitation. The 0-RPM action keeps the prism stationary and introduces no additional image-plane motion. The source sequence, timestamps, threshold-event model, and evaluation pipeline otherwise remain unchanged. It therefore serves as the passive DVS counterpart within our active-sensing backend and should not be confused with the separate native implementations of DVS-Voltmeter or v2e.

D.3 Privileged Teacher and Policy Training

The policy is trained by imitation from offline privileged supervision rather than reinforcement learning. For each training window i, we obtain image-wide edge-F1 scores from matched fixed-excitation sequences,

$$F_{1,i}(a), \qquad a \in \{0, 180, 360\}\ \mathrm{RPM}. \tag{11}$$

Each candidate sequence is generated open-loop at a fixed prism speed, and the teacher compares the same temporal window across the three candidates. The classification target is

$$a_i^\star = \arg\max_{a \in \mathcal{A}_p} F_{1,i}(a), \tag{12}$$

with ties resolved in favor of the lower excitation level. This supervision is a same-window comparison of fixed actions; it is not an on-policy rollout, future-horizon return, or TEO-based objective. TEO is used only for evaluation.

The teacher further incorporates causal excitation regularization based on the previous teacher-executed state. When the recent aligned-event support already indicates saturated or highly repetitive activity, the highest excitation candidate can be excluded to avoid assigning labels that favor unnecessary event redundancy. These regularization rules depend only on completed windows and do not use future observations.

The student is optimized with inverse-frequency weighted cross-entropy,

$$\mathcal{L}_{\text{cls}} = -\frac{1}{B}\sum_{i=1}^{B} w_{a_i^\star} \log p_\theta(a_i^\star \mid \mathbf{S}_i), \tag{13}$$

using AdamW with learning rate 2×10^{-3} and weight decay 10^{-4}. Training uses full-batch updates without a learning-rate scheduler, and the checkpoint with the lowest validation cross-entropy is retained.

At inference time, only completed sensing history is available. Before eight history windows have accumulated, the implementation defaults to the nominal 180-RPM action. Additional confidence, event-storm, and dwell safeguards prevent unstable rapid switching and reject unnecessarily strong excitation when recent aligned-event activity is already excessive.

Controlled policy-ablation instance. The general EIRM formulation uses the three-action space $\{0, 180, 360\}$ RPM. For the controlled optical-policy ablation reported in the main paper, the evaluated checkpoint uses the restricted two-action set $\{180, 360\}$ RPM. This restriction applies only to that ablation protocol and does not change the general EIRM action space.

D.4 Event-Driven Camera-Motion Feedback

The camera-motion feedback path is separate from the learned optical-excitation policy and is not learned jointly with it. It regulates progression along the controlled camera trajectory using aligned-event occupancy and its first-order temporal change.

Let $\mathcal{S}_n^{\text{ref}}$ denote the unique aligned-event support in the completed control window I_n, and let Ω_v denote the valid image region. We define

$$\rho_n = \frac{\left|\mathcal{S}_n^{\text{ref}}\right|}{|\Omega_v|}, \qquad \Delta\rho_n = \rho_n - \rho_{n-1}. \tag{14}$$

The rule-based command for the following interval is

$$u_{n+1}^m = \begin{cases} \text{Decelerate}, & \rho_n \geq 0.50 \text{ or } \big(\rho_n \geq 0.35 \text{ and } \Delta\rho_n \geq 0.03\big), \\ \text{Accelerate}, & \rho_n \leq 0.35 \text{ and } \Delta\rho_n \leq 0, \\ \text{Keep}, & \text{otherwise}. \end{cases} \tag{15}$$

The absolute occupancy detects sustained high event activity, while $\Delta\rho_n$ allows the controller to respond when activity is already rising within the intermediate occupancy range.

The command moves one adjacent slot among trajectory-progression rates

$$r_n \in \{0.5, 1.0, 1.5\} \times . \tag{16}$$

These values scale progression along the pre-recorded RGB trajectory; they do not directly specify a new physical angular or translational camera velocity. Control windows are evaluated at $1/60$ s, while a 0.20-s dwell constraint prevents rapid switching between rate levels. The rate never falls below $0.5\times$, and progression is truncated at the end of the available trajectory.

The optical and camera-motion branches therefore regulate complementary sources of sensing motion but do not share a learned encoder or a joint training objective.

Table 6: **Numerical consistency across event-generation backends.** All implementations use the same deterministic log-threshold event model. The reported event counts are obtained from a 960×540 sequence with 32 frames.

Backend	Total Events	Positive Events	Negative Events
CPU NumPy	928,532	442,923	485,609
PyTorch CUDA	928,532	442,923	485,609
Compiled CUDA	928,532	442,923	485,609

D.5 Event Generation and Timestamp-Aware Alignment

All optically modulated variants use the same temporal sampling, calibrated optical mapping, threshold-event generation, and timestamp-aware inverse mapping. For every generated event at timestamp t_e, the inverse optical warp is evaluated using the corresponding prism angle $\theta(t_e)$ rather than a representative interval pose.

This inverse warp compensates only prism-induced optical motion. Natural camera motion, object motion, and their relative image motion are intentionally preserved in the aligned event stream. Passive simulator baselines retain their native event coordinates. Unless otherwise stated, downstream evaluation of optically modulated variants uses the aligned reference-plane events after timestamp-aware compensation.

D.6 CUDA-Accelerated Event Simulation and Computational Efficiency

The proposed active-sensing pipeline requires repeated event regeneration under different optical-excitation policies, prism trajectories, and motion conditions. To enable efficient large-scale evaluation and policy analysis, we implement the computationally intensive components of EIRM with GPU acceleration, including prism warping, threshold-event generation, and timestamp-aware compensation.

GPU-accelerated components. The computational pipeline consists of three major GPU operators:

- **Prism warp.** The sensor-plane coordinates are mapped through the calibrated inverse prism model and resampled using bilinear interpolation. The operation is implemented with GPU tensor operations and performs image-space optical modulation before event generation.
- **Threshold-event generation.** Event generation follows the same deterministic log-intensity threshold model across CPU NumPy, PyTorch CUDA, and compiled CUDA implementations. The CUDA implementation parallelizes per-pixel temporal processing and event-buffer construction on the GPU. The compiled CUDA backend adopts a two-pass strategy: the first pass counts generated events and the second pass writes events into a preallocated buffer without atomic writes.
- **Timestamp-aware compensation.** For each generated event, the corresponding prism state is evaluated at its exact timestamp. Events are processed in GPU batches, and the inverse prism mapping is applied to recover the reference-plane coordinates.

The main experiments use the PyTorch CUDA backend because it supports the complete closed-loop sensing interface. The compiled CUDA extension provides an additional optimized implementation for standalone event-generation evaluation.

Implementation consistency. To verify numerical consistency, we compare CPU NumPy, PyTorch CUDA, and compiled CUDA implementations under the same deterministic threshold-event model. As shown in Table 6, all implementations produce identical event statistics on the same input sequence.

Throughput evaluation. We evaluate computational efficiency on a single NVIDIA RTX 4090 GPU under a matched 960×540, 32-frame, 60 Hz configuration. The compared methods receive

Table 7: **Computational throughput comparison.** Measurements are obtained under the matched event-generation protocol. Speedups are computed relative to each corresponding baseline runtime.

Method	Runtime (s) ↓	FPS ↑	Speedup
DVS-Voltmeter (CPU)	2.972	10.8	1.0×
v2e (CPU)	0.680	47.1	1.0×
AMI CPU threshold	3.232	9.9	1.0×
AMI PyTorch CUDA	**0.207**	**154.6**	**14.4× / 3.3× / 15.6×**

Table 8: **Runtime breakdown of the GPU sensing pipeline.**

Component	Runtime (s)
Prism warp	0.013
Threshold-event generation	0.206
Timestamp-aware compensation	0.224
Total	0.443

the same prism-warped sensor intensity sequence and timestamps. Data loading and controller execution are excluded from all measurements.

The GPU acceleration enables efficient repeated event regeneration required by active sensing. Compared with CPU implementations, the CUDA event-generation backend substantially reduces the computational cost while preserving identical event statistics. The remaining runtime is mainly attributed to timestamp-aware compensation, which requires evaluating the inverse optical mapping at individual event timestamps.

Temporal densification for DL3DV. The DL3DV rendering pipeline requires temporally dense visual signals before event regeneration. The rendered RGB sequences contain 40 frames per clip and are temporally densified to 480 frames using Practical-RIFE (Huang et al., 2022). Intermediate timestamps are assigned by linear interpolation between adjacent rendered frames. RIFE inference is performed with GPU acceleration before the generated sequences are passed to the same prism-warp, threshold-event generation, and timestamp-aware compensation pipeline.

Comparison with passive event simulators. The passive baselines use their native event-simulation implementations and do not share the active event-generation backend. DVS-Voltmeter models sensor-level effects including photoreceptor response and stochastic noise, whereas EIRM uses a deterministic calibrated log-threshold event model. v2e additionally models threshold mismatch and leakage effects. Therefore, event counts between different simulators are not expected to match and are not used as a quality criterion.

For fairness, passive baselines are compared using matched source RGB sequences, timestamps, temporal windows, and evaluation protocols. Fixed and adaptive active policies share the same prism-warp, threshold-event generation, timestamp-aware compensation, and evaluation pipeline.

E Evaluation Protocol and Metrics

E.1 Training, Validation, and Held-Out Evaluation

Teacher labels and feature-normalization statistics are constructed using the training split only. Model selection is performed on a separate validation split using validation cross-entropy. All quantitative results reported in the main paper are evaluated on held-out sequences that do not contribute to teacher construction, feature normalization, or checkpoint selection. The same fixed split assignment is used across compared active-policy variants.

Table 9: **Practical-RIFE temporal densification throughput on RTX 4090.** Measurements exclude model loading time.

Resolution	Runtime per interpolation	Source-interval FPS
960×540	92.3 ms	86.6
320×240	25.0 ms	320.3

Table 10: **Event simulator implementation comparison.**

Property	EIRM	DVS-Voltmeter	v2e
Backend	CUDA	CPU EventSim	EventEmulator
Event model	Log-threshold	Photoreceptor + noise	Threshold + leakage
Refractory	No	Model-dependent	Disabled
CUDA	Native	Not used	Partial
Timestamp	Event-level	Frame-level	Frame-level

E.2 Matched-Source Comparison

DVS-Voltmeter and v2e receive the same source RGB/luminance sequence and timestamps but retain their native simulator implementations. Fixed optical policies and EIRM instead use the same active prism-warp and threshold-event simulation path. Consequently, we do not assume a common event-generation backend across all methods.

Fairness is enforced through matched source trajectories, timestamps, temporal windows, and scoring procedures. Comparisons within the active-sensing family additionally hold the optical mapping, event simulator, temporal subdivision, and compensation pipeline fixed.

The controlled motion–excitation sweep evaluates fixed 0, 180, and 360 RPM actions. Fixed-180 serves as the nominal non-adaptive active baseline in the main F1 and TEO evaluation, while Fixed-360 provides a stronger constant-excitation setting in the policy and accumulation analyses. Fixed-0 uses the same active-sensing backend with the prism held stationary and therefore provides the matched zero-excitation condition.

E.3 Window-Level Aggregation

For the main edge evaluation, F1, precision, and recall are first computed for each temporal window and then arithmetically averaged. Mean F1 is therefore

$$\overline{F_1} = \frac{1}{N} \sum_{i=1}^{N} F_{1,i}, \tag{17}$$

and is not reconstructed from the separately averaged precision and recall. Regime-specific scores use the same evaluator after selecting only windows from the corresponding camera–object motion condition.

The main controlled split contains 12 clips across four motion conditions, with three seed-indexed clips per condition. Each condition contributes 189 evaluated windows, giving 756 windows per method over the full split.

E.4 Target Event Observability

For a temporal evaluation window, let T denote the benchmark target region, $G \subseteq T$ the target-edge set, and M the support of aligned events. The target edge is computed from the reference image inside T by applying a gradient operator to log-linear luminance and thresholding the gradient magnitude at the 75th percentile. The target region is obtained from the provided object-mask annotation and thresholded at 127.

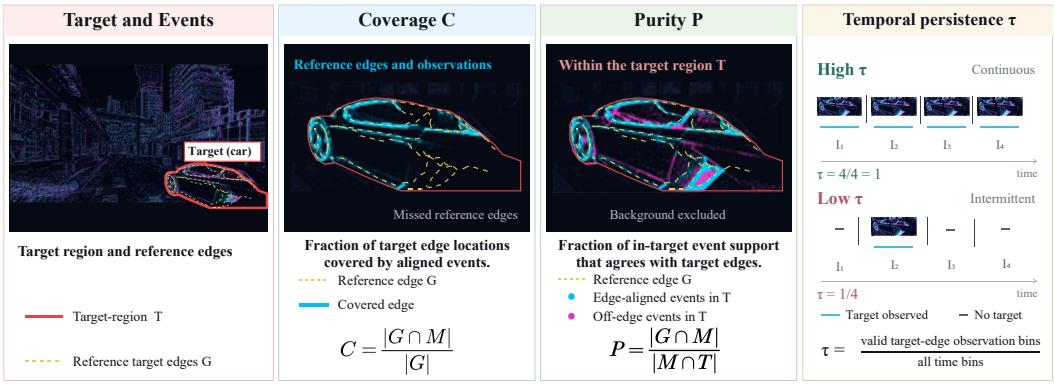


Figure 6: **Illustration of Target Event Observability.** Coverage C measures how much target-edge structure is observed, purity P measures how strongly target events concentrate on informative boundaries, and temporal persistence τ measures whether such observations remain present across the evaluated window. TEO combines the three factors to distinguish useful target observation from raw event abundance.

Target-edge coverage is

$$C = \frac{|G \cap M|}{|G|}, \tag{18}$$

and target purity is

$$P = \frac{|G \cap M|}{|M \cap T|}. \tag{19}$$

Background events do not enter the denominator of P. No dilation, matching radius, spatial tolerance, or minimum event-count threshold is used.

Each temporal window is divided into four equal bins. A bin is valid if it contains at least one target-edge event, giving

$$\tau = \frac{\#\{\text{valid temporal bins}\}}{4}. \tag{20}$$

The final metric is

$$\text{TEO} = \frac{2CP}{C+P}\tau. \tag{21}$$

If the required support is empty or $C + P = 0$, the corresponding score is set to zero. C, P, τ, and TEO therefore lie in $[0, 1]$. TEO is computed independently for each window and then arithmetically averaged over the selected split.

E.5 Aligned-Accumulation Fidelity

The reference used for FSIM and LPIPS is not a ground-truth event accumulation. Instead, we construct an RGB-derived structural reference from linear luminance Y_{lin} as

$$R_{\text{ref}}(\mathbf{x}) = \frac{\min(\|\nabla \log Y_{\text{lin}}(\mathbf{x})\|_2, q_{99.5})}{q_{99.5}}, \tag{22}$$

where $q_{99.5}$ denotes the shared train/validation 99.5th percentile of the log-luminance gradient magnitude.

For the event accumulation, positive and negative event counts are accumulated separately, followed by signed log compression and count clipping at 8 before mapping to $[0, 1]$. No per-method maximum normalization is used. For LPIPS, the scalar accumulation and reference maps are replicated to three channels and mapped to $[-1, 1]$; LPIPS uses the AlexNet backbone. FSIM uses the same normalized scalar maps with chromatic comparison disabled.

Both metrics are computed independently for each temporal window and then arithmetically averaged. The co-motion results use the identical protocol after filtering to the corresponding subset.

This experiment therefore evaluates the fidelity of aligned event-derived structure to an RGB-derived structural reference rather than to a hypothetical ground-truth event image.

F Additional Analysis of Adaptive Sensing

F.1 Motion–Excitation Response

The controlled sweep evaluates fixed $0/180/360$ RPM actions over camera-only, object-only, co-motion, and counter-motion conditions. Each plotted point in the main motion–excitation figure is the arithmetic mean over three seed-indexed rendered scene realizations. These realizations correspond to distinct scene instances rather than repeated stochastic reruns of the same render.

The sweep exposes two complementary quantities: target-level TEO and normalized event activity. Their separation is important because stronger optical motion can continue increasing event production after useful target observability has saturated. This behavior motivates adapting excitation to the current motion condition rather than selecting the strongest fixed action globally.

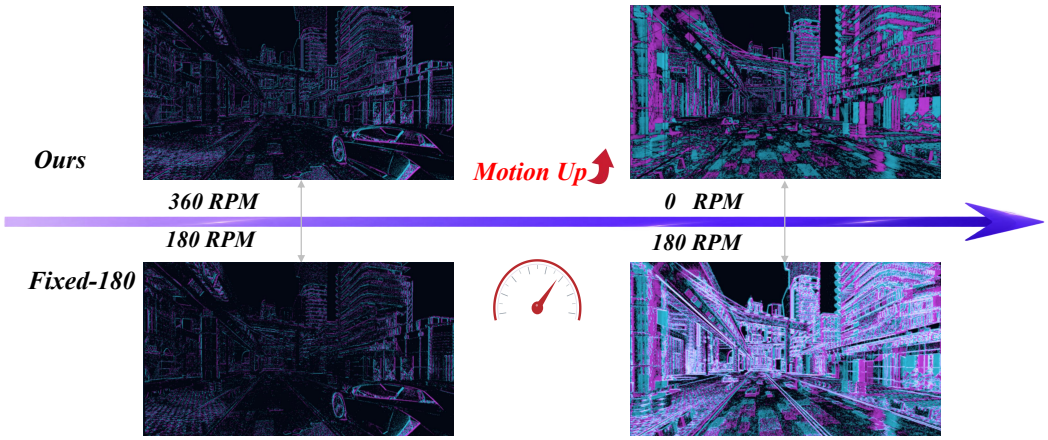


Figure 7: **Adaptive optical-excitation behavior.** Representative temporal behavior of the causal prism policy under changing sensing conditions. The policy uses completed sensing history to switch between the available excitation levels rather than maintaining a fixed prism speed throughout the sequence.

F.2 TEO Decomposition Analysis

To further analyze the factors contributing to Target Event Observability (TEO), we decompose the metric into its three components: target-edge coverage C, event purity P, and temporal persistence τ. We additionally report event density to distinguish useful target observation from raw event abundance. As shown in Table 11, EIRM improves TEO primarily through higher target-edge coverage while maintaining comparable purity and temporal persistence with Fixed-180. This indicates that the adaptive excitation improves the amount of observed target structure rather than merely increasing event activity.

F.3 Optical-Excitation Policy Diagnostics

The optical-policy diagnostics use the same controlled benchmark and evaluation protocol as the main experiments. We report Fixed-180 and Fixed-360 as the non-adaptive active-excitation baselines. Fixed-0 is included as the matched zero-excitation condition within the same active sensing backend.

For each sequence, window-level scores are first averaged within the sequence and the resulting sequence scores are then averaged across the evaluation set. EIRM operates at a mean actuation level of 197.1 RPM, while the stored mean event-density values are 0.4586 for Fixed-180, 0.6752

Table 11: **Decomposition of Target Event Observability (TEO).** We report target-edge coverage C, event purity P, temporal persistence τ, event density, and the 95% confidence intervals of TEO on the Blender evaluation subset.

Method	TEO ↑	C ↑	P ↑	τ ↑	Event density ↑	95% CI
DVS-Voltmeter	0.4891	0.3922	0.7008	0.7308	0.2724	[0.323, 0.646]
v2e	0.5328	0.4610	0.6978	0.7166	*0.4572*	[0.352, 0.705]
Fixed-180	*0.7728*	*0.6683*	**0.9442**	**1.0000**	0.4426	[0.757, 0.789]
EIRM (Ours)	**0.7772**	**0.6770**	*0.9433*	**1.0000**	**0.4704**	[0.761, 0.794]

for Fixed-360, and 0.4876 for EIRM. The adaptive policy can therefore reduce, maintain, or increase optical excitation according to the completed sensing history rather than remaining at a fixed prism speed.

These diagnostics show that the evaluated policy operates predominantly near the moderate-excitation regime while selectively invoking stronger excitation as the observed sensing state changes. This behavior is consistent with the motion-dependent response in the main paper, where the utility of additional optical excitation depends on the existing camera–object relative motion.

G Timestamp-Aware Compensation

The compensation ablation uses the same generated sensor-event stream for all three variants. The variants differ only in how the known prism trajectory is inverted:

- **No compensation**: events remain in sensor coordinates;
- **Interval-level**: one inverse optical warp is evaluated at the interval midpoint and applied to all corresponding events;
- **Per-event**: the inverse optical warp is evaluated independently at every event timestamp.

All three variants therefore contain identical generated events before alignment. The inverse warp removes only prism-induced optical motion and does not cancel natural camera or object motion.

The reported F1 and precision measure whole-image alignment to RGB-derived reference edges. The reported coverage,

$$\text{Cov.} = \frac{\#\{\text{active event pixels touching a reference edge}\}}{\#\{\text{reference-edge pixels}\}}, \tag{23}$$

is also a whole-image measure and is distinct from the target-level coverage C used by TEO.

The reconstructed-DL3DV compensation evaluation uses five scenes at width 160; the corresponding procedural diagnostic is restricted to a camera-only subset. Because the generated event stream is held fixed between the three variants, the measured differences isolate the temporal precision of the inverse optical warp.

H Evaluation on Recorded Visual–Inertial Motion

We use DSEC to examine the offline EIRM sensing pipeline under recorded RGB–IMU trajectories. This experiment should be interpreted as evaluation on real visual–inertial motion rather than as hardware validation of the rotating optical element.

The native DSEC event stream is retained only as an unaligned recorded reference. Passive, fixed-excitation, and EIRM event streams are generated offline from the DSEC RGB and IMU measurements using 50-ms temporal windows at 640×480 simulation resolution.

The stored diagnostics include event count, spatial coverage, raw and compensated flicker, gyroscope magnitude, and commanded prism speed. We use these sequences as a qualitative and descriptive check of EIRM under realistic visual–inertial dynamics rather than as a formal F1, TEO, FSIM, or LPIPS benchmark.

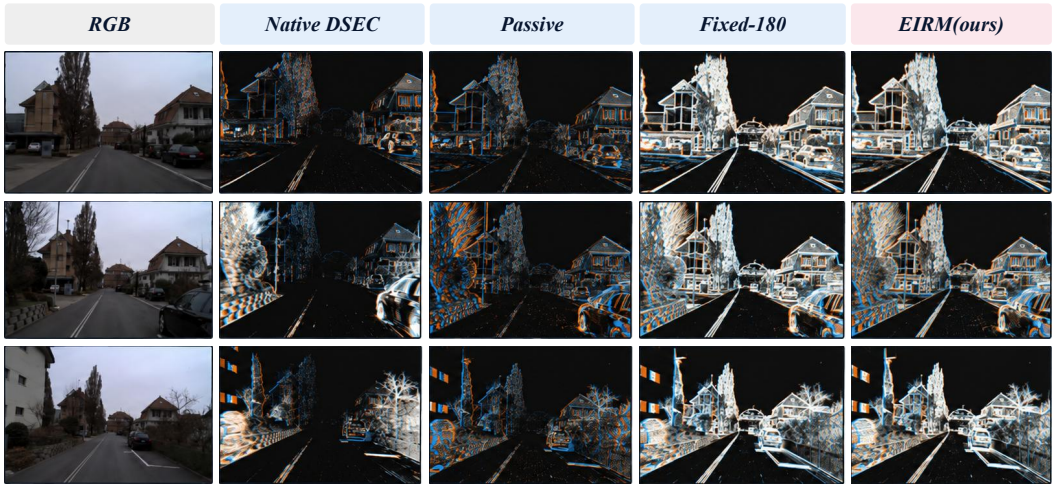


Figure 8: **Qualitative evaluation on recorded visual–inertial motion.** DSEC provides natural image motion and real inertial dynamics, while the compared sensing variants are generated offline from matched RGB–IMU observations. Native DSEC events are included only as a recorded reference; this experiment evaluates behavior under realistic motion statistics rather than hardware-level optical modulation.

I Additional Diagnostic and Consistency Checks

I.1 Why Passive Co-Motion Scores Can Be Zero

The zero passive scores under co-motion do not imply an empty event stream. The stored DVS-Voltmeter and v2e windows contain substantial background event activity but no events inside the evaluated target mask. Representative DVS-Voltmeter windows contain 312, 1353, and 1429 total events, respectively, while their target-event counts are all zero. Representative v2e windows contain 1033, 2192, and 2157 total events, again with zero target events.

Consequently, these windows have no overlap with the evaluated target edges: $C = 0$, $P = 0$, and $\tau = 0$, giving $\mathrm{TEO} = 0$. Their target-edge F1 is also zero because the evaluated overlap contains no true positives. Passive co-motion sequences may therefore retain visible background activity while providing no valid target-level observation.

I.2 Protocol Consistency and Numerical Checks

Matched inputs. Within each experiment, compared methods use matched source trajectories, timestamps, temporal windows, and evaluation timestamps. Passive simulator backends remain distinct, as described in Sec. E.

Active-policy consistency. Fixed and adaptive optical policies use the same calibrated optical mapping, threshold-event simulator, temporal subdivision, and compensation pipeline, with the excitation policy as the controlled variable.

Policy-split consistency. Teacher construction and feature normalization use training data only, model selection uses the validation split, and held-out evaluation sequences are not used for either purpose.

Motion-feedback consistency. Camera-motion feedback uses only aligned-event observations from completed windows. Its occupancy thresholds, trajectory-rate levels, and dwell constraint are fixed across evaluated sequences, and no future event information is used to determine the next motion command.

Compensation consistency. No-, interval-level, and per-event compensation variants share the identical generated event stream. Their differences therefore isolate inverse-warp timing.

Metric consistency. Target masks, RGB-derived reference edges, valid image support, temporal windows, and aggregation rules are fixed within each reported experiment.