EVMARS-HARSH: AN RGB-EVENT MULTIMODAL BENCHMARK FOR ROBUST DETECTION IN MARTIAN DUST STORM ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Event cameras have recently attracted increasing attention in computer vision due to their high temporal resolution, high dynamic range, and low power consumption. However, existing event-based detection benchmarks mainly focus on indoor or real-world driving scenarios, while perception degradation under extreme environments remains insufficiently explored. In planetary exploration and autonomous inspection, severe dust disturbances and complex illumination conditions can degrade visual observations and hinder the spatio-temporal characteristics of event streams, making robust detection challenging. To this end, we introduce EVMars-Harsh, an RGB-Event multimodal detection benchmark for harsh dust environments. EVMars-Harsh is collected in a Mars-like experimental field using a mobile rover platform with synchronized RGB and Color Event sensors, where a controllable blower system is employed to simulate Martian dust disturbances and severe perception degradation. Based on EVMars-Harsh, considering the reduced reliability of event streams under dust interference, we propose a Dust-Aware Sampling module (DAS), which establishes spatio-temporal correlations through scale-adaptive structural continuity analysis to suppress noise events caused by dust particles. Then, we introduce a Reliability-Aware Multimodal Fusion (RAMF) with spatial reliability estimation to adaptively integrate RGB and Event features. Moreover, a Reliability-Aware Lightweight Axial Selective State Space Module (RA-LASSM) is designed to capture long-range dependencies while reducing noise propagation in degraded regions. Experiments on EVMars-Harsh show that the proposed framework achieves 0.743 mAP while maintaining low model complexity, demonstrating its effectiveness for robust RGB-Event detection under severe dust conditions.

1 INTRODUCTION

Event cameras offer high temporal resolution, high dynamic range, and low data redundancy Hao et al. (2026), showing strong potential in autonomous driving, robotic perception, and space exploration. By asynchronously responding to pixel-level brightness changes, they efficiently capture fast motion and local dynamics, while conventional frame cameras provide rich color, texture, and semantic information Lu et al. (2025); Yu et al. (2026). Their complementary characteristics are particularly valuable for object detection in extreme environments, where a single modality often fails to preserve both appearance and dynamic cues. Therefore, establishing RGB-Event multimodal detection datasets is important for advancing robust visual perception.

In recent years, various event-based vision datasets have been developed for object detection, autonomous driving, and robotic perception. Existing datasets are mainly collected in indoor environments, urban roads, or natural driving scenarios, with challenges typically involving high-speed motion, dynamic backgrounds, day-night variations, and high-dynamic-range conditions associated with tunnel entry and exit Gehrig et al. (2021); Perot et al. (2020); Li et al. (2023). These datasets have provided important support for event-based vision research, yet systematic investigation of scene perception degradation caused by extreme environmental factors remains limited Ma et al. (2026). Existing event-based detection datasets are therefore insufficient to fully characterize the practical perception difficulties encountered by autonomous exploration systems operating in extreme and unstructured environments. This issue is particularly pronounced in planetary rover mis-

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.

sions. Mars rovers are required to perform rock recognition, terrain understanding, and autonomous navigation in environments lacking structured roads and stable visual conditions. Their perception systems may simultaneously encounter severe dust interference, intense illumination changes, local overexposure, object occlusion, and complex and uneven terrain. These factors not only significantly degrade RGB observations but also alter the spatio-temporal response distribution of event streams.

To deal with these challenges, we construct EVMars-Harsh, an RGB-Event multimodal object detection benchmark in Mars-like dust environments. Unlike existing datasets mainly collected in road-driving or indoor scenarios, EVMars-Harsh is acquired using a mobile rover in a constructed Mars-like test field containing deep pits, shallow pits, slopes, and rough terrain. During data collection, the sensors face multiple extreme environmental challenges, including dust disturbances actively generated by a blower system, intense illumination, and local overexposure. Consequently, the degradation is caused by the combined effects of real environmental factors during data acquisition, rather than by artificially synthesized degradation in post-processing. EVMars-Harsh simultaneously records RGB images and color event data. The color event stream preserves spatial, temporal, polarity, and color information, enabling complementary Color Event and Polarity Event representations. Its challenges arise not only from variations in object scale, color, and texture, but also from persistent perception degradation. Severe dust may occlude objects and generate numerous random events; strong overexposure may weaken RGB appearance cues; and pits and slopes can continuously alter viewing angles, object scales, and motion patterns. Therefore, the core problem studied by EVMars-Harsh is not conventional object detection under normal conditions, but rather whether RGB and event modalities can maintain reliable and complementary object perception when visual information is subject to persistent and compound environmental degradation.

Along with EVMars-Harsh, we propose a reliability-aware multimodal benchmark framework for robust RGB-Event detection in extreme dust environments, targeting three key challenges: event stream contamination, spatially varying modality reliability, and long-range structural modeling under degraded conditions. First, to address the large number of random events caused by dust particles, we propose a Dust-Aware Sampling (DAS) module, which leverages the spatio-temporal continuity and multi-scale structural stability of events to retain informative events with stable object structures while suppressing random dust-induced responses. Second, to address the dynamically varying reliability of different modalities under degraded conditions, we design a Reliability-Aware Multimodal Fusion (RAMF) module, which adaptively adjusts the contributions of RGB, Color Event, and Polarity Event features according to spatial locations while estimating local reliability. Finally, we introduce a Reliability-Aware Lightweight Axial Selective State-Space Module (RA-LASSM), which employs axial state-space modeling to capture long-range spatial dependencies while leveraging reliability information to suppress noise propagation from degraded regions. These designs collaborate closely to address the extreme environment perception challenges presented by EVMars-Harsh at three levels: event input quality, multimodal information selection, and deep feature propagation.

Under a unified evaluation protocol, EVMars-Harsh enables systematic comparison of different visual modalities and object detection methods. Beyond providing a new RGB-Color Event multimodal dataset, it establishes a benchmark for studying scene perception degradation jointly caused by severe dust, overexposure, occlusion, and complex terrain, providing an experimental foundation for robust multimodal perception in robotics in extreme environments and planetary exploration. The main contributions are summarized as follows:

- **EVMars-Harsh Dataset.** We construct EVMars-Harsh, an RGB-Event multimodal detection benchmark for Martian dust storm environments. It is collected by a mobile rover in a manually constructed Mars-like field containing severe dust, intense illumination and overexposure, object occlusion, and complex unstructured terrain, which extends event-based object detection to perception degradation under extreme environmental conditions.
- **Dust-Aware Sampling.** To address numerous random event responses caused by dust particles, we propose DAS, which suppresses random dust events through spatio-temporal continuity and scale-adaptive structural stability analysis while preserving effective structural information of rock objects at different scales, improving the robustness of event representation in harsh environments.
- **Reliability-aware Multimodal Detection Framework.** We propose a reliability-aware RGB-Event detection framework for extreme dust environments. RAMF adaptively bal-

ances RGB, Color Event, and Polarity Event representations based on spatially varying reliability, while RA-LASSM injects reliability cues into axial state-space modeling to suppress the propagation of unreliable features. Together with DAS, they form a unified framework addressing event input quality, multimodal information selection, and reliable long-range feature modeling.

2 Related Works

With the widespread application of event cameras in object detection, various event-based datasets have been developed for different scenarios Sun et al. (2024); Zhou et al. (2024). Existing benchmarks can be broadly categorized into event-only and RGB-Event multimodal detection datasets.

Event-only Detection Datasets. Early event-based detection datasets mainly focused on autonomous driving scenarios. The Gen1 Automotive Detection Dataset was collected in urban, suburban, and highway environments and provides bounding-box annotations for vehicles and pedestrians, serving as an important benchmark for event-based object detection de Tournemire et al. (2020). The later 1Mpx Automotive Detection Dataset increased the sensor resolution to the megapixel level and contains more than 14 hours of driving data Perot et al. (2020). Although grayscale frames are also provided, most studies mainly rely on event streams. Recent datasets further extend event-based detection to more diverse scenarios. eTraM targets fixed-view traffic monitoring and contains about 10 hours of event data with nearly two million bounding boxes under day–night changes, rain, glare, and exposure variations Verma et al. (2024). EvDET200K includes 10 categories, about 200,000 bounding boxes, and 10,054 samples, improving category diversity for generic event-based detection Wang et al. (2025). EV-UAV focuses on tiny UAV detection, with 147 event sequences and more than 2.3 million event-level annotations Chen et al. (2025b).

RGB-Event multimodal detection datasets. Beyond event-only detection, the complementary characteristics of RGB images and event streams have also motivated the development of multimodal detection datasets. PKU-DAVIS-SOD uses a DAVIS346 sensor to synchronously capture RGB frames and events, providing 220 sequences and more than 1.08 million bounding boxes under conditions such as normal illumination, low light, and high-speed motion Li et al. (2023). It has become an important benchmark for continuous RGB-Event detection. DSEC was originally introduced for stereo vision and optical flow in autonomous driving, with synchronized color frames and event data Gehrig et al. (2021). Based on DSEC, subsequent works released or derived detection benchmarks including DSEC-Detection, DSEC-Det, and DSEC-MOD for RGB-Event object and moving-object detection Gehrig & Scaramuzza (2024); Liu et al. (2024); Zhou et al. (2023).

Overall, existing event-based detection datasets cover autonomous driving, traffic monitoring, tiny objects, and low-light conditions. However, their challenges mainly focus on motion, illumination changes, and complex traffic backgrounds, while persistent perception degradation caused by multiple extreme environmental factors remains underexplored. Moreover, most public datasets rely on monochromatic event streams, leaving color event, polarity event, and their complementarity with RGB under extreme conditions underexplored. To address this gap, we construct EVMars-Harsh, extending event-based detection to Mars-like harsh environments with severe dust, strong exposure, object occlusion, and complex unstructured terrain during real data acquisition.

3 EVMars-Harsh Benchmark

To address the limited coverage of extreme environmental degradation in existing event-based detection datasets, we construct EVMars-Harsh, an RGB-Event multimodal detection benchmark in Mars-like dust environments. As shown in Fig. 1, data are collected using a mobile Mars-like rover under physical disturbances, causing simultaneous degradation of RGB appearance and event spatio-temporal responses.

Category Taxonomy and Task Design. Existing planetary event-based datasets mainly target SLAM, navigation, or visual odometry (Mahlknecht et al., 2022), and are not directly suited to object detection. EVMars-Harsh defines eight rock classes: Weathered Basalt, Mudstone, and Normal, Obstacle, and Shelter variants of Basalt and Sandstone. The taxonomy jointly considers geological representativeness and rover perception requirements. Basalt, Weathered Basalt, Sandstone, and Mudstone represent different lithological and visual characteristics (McSween et al., 2004; McSween Jr, 2015; Wyatt & McSween Jr, 2002; Hausrath et al., 2008; Vaniman et al., 2014; Yen et al., 2017). For structurally stable Basalt and Sandstone, we introduce Normal, Obstacle, and Shelter

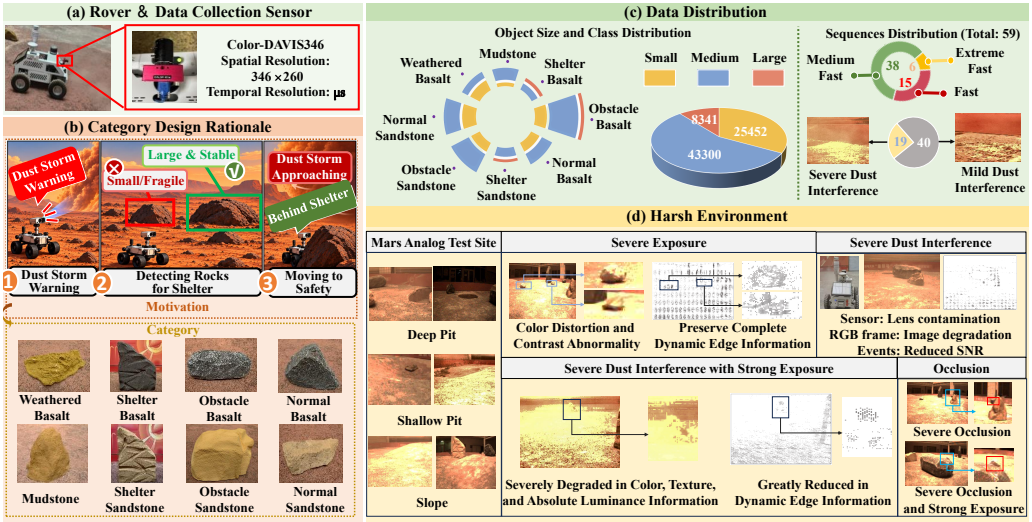


Figure 1: **EVMars-Harsh Benchmark.** The data are collected using a mobile rover, with rock categories designed according to Mars exploration requirements. Object and sequence distributions are also analyzed, while the dataset covers representative extreme conditions including severe dust, overexposure, object occlusion, and complex terrain.

categories according to object scale and functional roles in rover navigation. This design enables evaluation of both fine-grained lithological discrimination and the detection of rocks with different scales and task-oriented attributes, as shown in Fig. 1(b).

Data Annotation. All sensors are hardware-synchronized using unified timestamps, enabling microsecond-level alignment between APS frames and Color Event streams. Object annotations are manually generated in YOLO format as $(c_j, x_j, y_j, w_j, h_j)$, where c_j denotes the class label, (x_j, y_j) the normalized bounding-box center, and w_j, h_j the normalized width and height.

Construction of the Mars-like Harsh Environment. The primary objective of EVMars-Harsh is to capture the degradation of compound perception caused by real environmental factors. As shown in Fig. 1(d), the Mars-like test field contains unstructured terrain such as deep pits, shallow pits, slopes, and rough surfaces, introducing variations in rover pose, viewpoint, and object scale. A blower system is employed to generate dust disturbances during data acquisition. Dust reduces RGB contrast, color fidelity, texture visibility, and object completeness, while moving particles trigger additional non-target event responses. The dataset also includes strong illumination, local overexposure, and varying degrees of object occlusion. In particular, severe dust combined with strong exposure can simultaneously degrade both RGB and event observations through different mechanisms. Therefore, EVMars-Harsh focuses on robust multimodal perception under the combined effects of severe dust interference, strong exposure, object occlusion, and complex terrain variations.

Dataset Statistics and Distribution. EVMars-Harsh contains 59 acquisition sequences and approximately 440 million color events. The average sequence duration is 23.77 s, ranging from 5.20 s to 53.97 s. Each sequence contains approximately 7.486×10^6 events on average, corresponding to an average event rate of 3.149×10^5 events/s. As shown in Fig. 1(c), the dataset covers different rover motion speeds and dust degradation levels. Mild and severe dust interference sequences account for 67.80% and 32.20%, respectively. The dataset contains 77,093 object instances, where Small, Medium, and Large objects account for 33.01%, 56.17%, and 10.82%, respectively.

Evaluation Protocol. For fair comparison, all methods are evaluated using the same data split and training settings following the COCO object detection protocol. We adopt $\mathrm{mAP}_{50:95}$ as the primary metric to jointly evaluate classification and localization performance.

4 Methodology

Overview. As shown in Fig. 2, we propose a reliability-aware RGB-Event multimodal detection framework for robust detection in Martian dust storm environments. First, DAS suppresses dust-

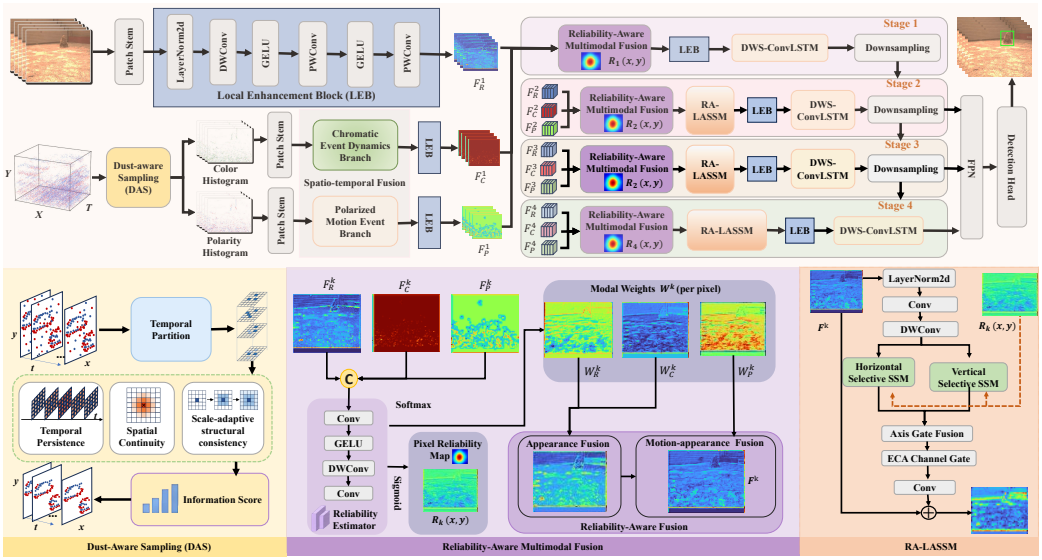


Figure 2: **The Overview of the Proposed Reliability-Aware Multimodal Detection Benchmark.** DAS suppresses dust-induced event noise and constructs Color Event and Polarity Event representations. RAMF adaptively fuses RGB and event features according to local reliability, while RA-LASSM performs reliability-guided long-range feature modeling. Multi-stage temporal aggregation and multi-scale fusion are finally used for object detection.

induced random responses at the raw-event level using spatio-temporal continuity and scale-adaptive structural consistency, from which Color Event and Polarity Event representations are constructed. The RGB, Color Event, and Polarity Event branches then extract complementary appearance and dynamic features. At each stage, RAMF estimates pixel-wise modality contributions and spatial reliability to perform progressive multimodal fusion. Subsequently, RA-LASSM incorporates the estimated reliability into axial state-space modeling to suppress long-range propagation of unreliable features. Finally, multi-stage temporal aggregation and multi-scale feature fusion are employed for object classification and localization.

4.1 Dust-aware Event Sampling

Under severe dust conditions, moving particles and sensor background activity can generate numerous non-target events, which are accumulated together with informative responses from rock edges and textures and consequently degrade event representations. To address this issue, as shown in Fig. 3, we propose a Dust-aware Event Sampling (DAS) module that evaluates and adaptively samples raw events before constructing event representations. Given a raw color event stream $\mathcal{E} = \{e_i = (x_i, y_i, t_i, p_i, c_i)\}_{i=1}^{N}$, we first divide the temporal window $[t_0, t_1]$ into M micro-temporal intervals and construct binary occupancy maps:

$$\begin{cases} m_i = \text{clip}\left(\left\lfloor M\dfrac{t_i - t_0}{t_1 - t_0}\right\rfloor, 0, M-1\right), \\ O_m(x,y) = \mathbb{I}\left[\exists e_i : m_i = m, (x_i, y_i) = (x, y)\right]. \end{cases} \tag{1}$$

Binary occupancy reduces the influence of repeated activations at the same location on event-quality estimation. The spatio-temporal continuity of each event is then jointly evaluated using spatial neighborhood support and temporal persistence:

$$\begin{cases} A_i = \displaystyle\sum_{(u,v)\neq(x_i,y_i)} G_\sigma(u - x_i, v - y_i) O_{m_i}(u, v), \\ T_i = \dfrac{1}{Z_i}\displaystyle\sum_{d\in\mathcal{D}_i} \dfrac{1}{|d|}\widetilde{A}_{m_i+d}(x_i, y_i), \\ C_i = \sqrt{A_i T_i}, \end{cases} \tag{2}$$

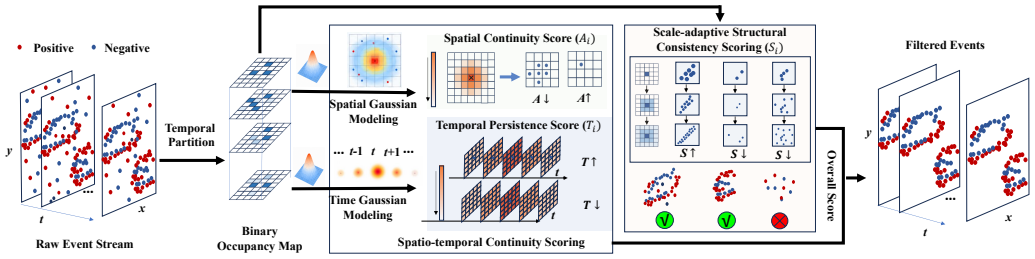


Figure 3: the proposed Dust-Aware Sampling (DAS). DAS jointly evaluates spatio-temporal continuity and scale-adaptive structural consistency to suppress dust-induced random responses while preserving informative events with stable object structures.

where G_σ denotes a Gaussian spatial kernel, $\mathcal{D}_i$ is the set of valid temporal offsets, and Z_i is the corresponding normalization term. A high C_i therefore indicates both stable spatial support and temporal persistence. To accommodate rock structures of different sizes, DAS introduces scale-adaptive structural consistency. For the local event support U_i^k at spatial scale k, the support growth between adjacent scales is combined with local directional coherence:

$$\begin{cases} U_i^k = \displaystyle\sum_{(u,v)\in\Omega_k(x_i,y_i)} O_{m_i}(u,v), \\ G_i^{(j)} = \mathrm{clip}\left(\dfrac{U_i^{k_{j+1}} - U_i^{k_j}}{U_i^{k_{j+1}} + \epsilon}, 0, 1\right), \\ L_i^k = \dfrac{\lambda_1^k - \lambda_2^k}{\lambda_1^k + \lambda_2^k + \epsilon}, \qquad S_i = \max\limits_j \sqrt{G_i^{(j)} L_i^{(j)}}. \end{cases} \tag{3}$$

Here, $\lambda_1^k \geq \lambda_2^k$ are the eigenvalues of the local covariance matrix. $L_i^{(j)}$ denotes the local directional coherence evaluated at the larger scale k_{j+1} of the j-th adjacent scale pair. Rock boundaries tend to exhibit meaningful support growth and directional coherence within an appropriate spatial range, whereas randomly scattered dust responses generally lack such stable structures. The specific scale settings are provided in the experimental section. Finally, DAS adaptively combines continuity and structural information according to the statistics of the current event window and performs probabilistic sampling. Specifically, the continuity weight w_c and structural weight w_s are obtained by normalizing the mean responses of the two scores within the current window. The information score of the i-th event is defined as $I_i = w_c C_i + w_s S_i$, and the mean information score $\alpha = \frac{1}{N}\sum_{i=1}^{N} I_i$ is used as an adaptive sampling threshold. The retention probability is defined as

$$P_i = \begin{cases} 1, & I_i \geq \alpha, \\ \left(\dfrac{I_i}{\alpha}\right)^\gamma, & I_i < \alpha. \end{cases} \tag{4}$$

Informative events are directly retained, whereas lower-score events are sampled according to their relative information content. Since both the fusion weights and sampling threshold are adaptively determined from the current event window, the strategy does not require re-tuning for different dust intensities. Consequently, DAS suppresses dust-induced random responses while preserving structurally informative events, providing more reliable inputs for subsequent Color Event, Polarity Event, and multimodal fusion.

4.2 Reliability-Aware Multimodal Fusion

Even after DAS preprocessing, the reliability of RGB and Event modalities remains spatially varying under extreme environmental conditions. For example, dust and overexposure may degrade RGB appearance cues, while moving particles can introduce erroneous responses into event streams. To address this issue, we propose a **Reliability-Aware Multimodal Fusion (RAMF)** module, which jointly models the RGB representation and two complementary Event representations, i.e., Color Event and Polarity Event, and adaptively adjusts their contributions according to local observations.

Given the RGB feature F_R^k, Color Event feature F_C^k, and Polarity Event feature F_P^k at the k-th stage, we first concatenate them along the channel dimension and obtain a joint representation through a lightweight reliability encoder:

$$Z^k = \mathcal{E}_{\text{rel}}^k\left(\text{Cat}\left[F_R^k, F_C^k, F_P^k\right]\right). \tag{5}$$

Here, $\mathcal{E}_{\text{rel}}^k(\cdot)$ models the cross-modal interaction between RGB and Event information, as well as the complementarity between the Color Event and Polarity Event representations within the Event modality, producing the joint representation Z^k. Based on Z^k, RAMF predicts pixel-wise contribution weights for the three representations together with a local spatial reliability map:

$$\begin{cases} [W_R^k, W_C^k, W_P^k] = \text{Softmax}\left(f_w(Z^k)\right), \\ R_k(x,y) = \sigma\left(f_r(Z^k)\right). \end{cases} \tag{6}$$

Here, W_R^k, W_C^k, and W_P^k represent the relative contributions of the RGB, Color Event, and Polarity Event representations at each spatial location, respectively, while $R_k(x,y)$ measures the overall reliability of the local RGB-Event observation. Since RGB and Color Event mainly provide appearance- and color-related information, whereas Polarity Event emphasizes brightness variations and dynamic structures, RAMF adopts a progressive fusion strategy rather than directly combining the three representations in a single weighted sum. Specifically, the RGB and Color Event weights are first renormalized to construct an appearance representation:

$$\widetilde{W}_R^k = \frac{W_R^k}{W_R^k + W_C^k + \epsilon}, \qquad \widetilde{W}_C^k = \frac{W_C^k}{W_R^k + W_C^k + \epsilon}, \tag{7}$$

$$F_A^k = \phi_A\left(\widetilde{W}_R^k F_R^k + \widetilde{W}_C^k F_C^k\right). \tag{8}$$

The Polarity Event representation is then incorporated to provide complementary dynamic information:

$$F_F^k = \phi_M\left[(1 - W_P^k)F_A^k + W_P^k F_P^k\right]. \tag{9}$$

Here, $\phi_A(\cdot)$ and $\phi_M(\cdot)$ denote lightweight local refinement operations. This progressive strategy first integrates the appearance information provided by RGB and Color Event, and then introduces the Polarity Event representation to complement dynamic structural cues, thereby enabling adaptive RGB-Event fusion. Meanwhile, the predicted reliability map $R_k(x,y)$ is further used by RA-LASSM to guide reliability-aware long-range feature propagation.

4.3 Reliability-Aware Axial Selective State-Space Module

Although RAMF adaptively adjusts the contributions of different modalities according to local degradation, unreliable responses in the fused representation may still propagate over long spatial ranges. To address this issue, we propose a Reliability-Aware Lightweight Axial Selective State-Space Module (RA-LASSM), which explicitly incorporates spatial reliability into axial state-space modeling to suppress the propagation of unreliable information.

Given the fused feature F_F^k and reliability map R_k at the k-th stage, a lightweight projection first produces a compact representation X, which is then serialized along the horizontal and vertical axes. For each direction $a \in \{h, v\}$, bidirectional state-space modeling is formulated as

$$Y_a = \mathcal{U}_a^{-1}\left[\frac{1}{2}\left(\mathcal{S}_a(X_a, R_a) + \text{Rev}\left(\mathcal{S}_a(\text{Rev}(X_a), \text{Rev}(R_a))\right)\right)\right], \tag{10}$$

where $(X_a, R_a) = \mathcal{U}_a(X, R_k)$. Here, $\mathcal{U}_a(\cdot)$ denotes serialization of the feature and reliability map along direction a, while $\mathcal{U}_a^{-1}(\cdot)$ restores the spatial layout. $\mathcal{S}_a(\cdot)$ denotes the Selective SSM for the corresponding direction, with independent parameters for horizontal and vertical modeling. $\text{Rev}(\cdot)$ denotes sequence reversal along the current scanning axis, and the outer reversal restores the backward output to the original spatial order. The forward and backward scans within each axis share the same parameters, enabling bidirectional long-range dependency modeling. The key mechanism of RA-LASSM is reliability-guided state writing. For each sequence position t, the spatial reliability R_t modulates the strength of writing the current input into the state:

$$\widetilde{B}_t = B_t \left[r_{\text{floor}} + (1 - r_{\text{floor}}) R_t \right], \qquad h_t = d_t h_{t-1} + (1 - d_t) \widetilde{B}_t u_t, \tag{11}$$

where u_t denotes the input feature at position t, h_t is the current state, B_t is the input-dependent state-writing coefficient, and d_t is the input-dependent state-decay factor. $R_t \in [0, 1]$ denotes the spatial reliability, and r_{floor} is a minimum reliability factor. Therefore, reliable features are sufficiently written into the state, whereas low-reliability responses are suppressed before long-range propagation. The non-zero r_{floor} further preserves a minimum information path for potentially informative weak responses. After obtaining the horizontal and vertical representations Y_h and Y_v, their contributions are adaptively determined according to the local feature:

$$[\alpha_h, \alpha_v] = \text{Softmax}\left(f_{\text{axis}}(X) \right), \qquad Y = \alpha_h \odot Y_h + \alpha_v \odot Y_v, \tag{12}$$

where $f_{\text{axis}}(\cdot)$ predicts the axial weights, α_h and α_v denote the spatially varying contributions of the horizontal and vertical directions, respectively, and $\odot$ denotes element-wise multiplication. The fused representation is finally refined by lightweight channel reweighting and residual restoration. By combining reliability-guided state writing with bidirectional axial modeling, RA-LASSM enhances long-range dependencies of reliable structures while suppressing the spatial propagation of degraded responses caused by dust and overexposure.

5 Experiments

Experimental Settings. Object detection performance is evaluated using the COCO-style average precision metric, with $\text{mAP}_{50:95}$ adopted as the primary evaluation metric. To construct the event representations, we use a fixed event window of $50\,\text{ms}$ and uniformly divide each window into $T = 10$ temporal slices. For the proposed DAS, the scale-adaptive structural modeling adopts the spatial scale set $K = \{3, 7, 15, 31\}$. The model is trained end-to-end using the AdamW optimizer, with the maximum learning rate set to 10^{-4} and the weight decay set to 0, together with the OneCycleLR learning rate scheduling strategy. During inference, the confidence threshold and NMS threshold are fixed to 0.1 and 0.45, respectively. We adopt the YOLOX detection head Ge et al. (2021) and use the standard YOLOX detection loss for training.

5.1 Ablation Studies

To analyze the contributions of different input modalities and individual components, we conduct all ablation experiments using the same data split, training strategy, and evaluation protocol. Except for the input modality or functional module under investigation, all other settings are kept unchanged.

Table 1: Ablation study on different input representations.

Setting	RGB	Color Event	Polarity Event	mAP ↑
RGB-only	✓	-	-	0.657
Event-only	-	✓	✓	0.681
Polarity Event-RGB	✓	-	✓	0.712
Color Event-RGB	✓	✓	-	0.720
Event-RGB	✓	✓	✓	**0.743**

Effect of input representations. All ablation experiments are conducted using the full network configuration, with only the available input representations varied. When a representation is removed, its corresponding feature branch is disabled, while the remaining applicable modules and detection head are kept unchanged. As shown in Tab. 1, incorporating event representations improves detection performance over RGB-only input, while jointly using RGB, Color Event, and Polarity Event achieves the best performance of 0.743 mAP. These results demonstrate the complementarity between RGB appearance cues and the chromatic and dynamic structural information provided by event representations.

Ablation of individual components. To evaluate the contribution of each component, we construct a basic multimodal detector as the baseline. It retains the RGB, Color Event, and Polarity Event encoders, temporal aggregation, and the YOLOX detection head, while removing DAS, RAMF, and RA-LASSM. Raw events are directly used to construct event representations, and fixed progressive

Table 2: Ablation study on different components.

Setting	DAS	RAMF	Axial SSM	Reliability	mAP ↑
Baseline	-	-	-	-	0.647
+ DAS	✓	-	-	-	0.675
+ RAMF	✓	✓	-	-	0.698
+ LASM	✓	✓	✓	-	0.719
Full	✓	✓	✓	✓	**0.743**
Full w/o DAS	-	✓	✓	✓	0.708

fusion is adopted instead of RAMF. Specifically, RGB and Color Event features are first fused with equal weights, followed by an equivalent uniform fusion with the Polarity Event feature. As shown in Tab. 2, the baseline achieves 0.647 mAP. Adding DAS, RAMF, and Axial SSM progressively improves the performance to 0.675, 0.698, and 0.719, respectively, demonstrating the benefits of event denoising, adaptive fusion, and long-range dependency modeling. Incorporating reliability-guided state propagation in RA-LASSM increases the performance to 0.743 mAP. Removing DAS from the full model reduces the performance to 0.708, highlighting the importance of high-quality event inputs for subsequent reliability-aware fusion and long-range feature modeling.

5.2 Results

Table 3: Performance comparison with state-of-the-art methods.

Method	Modality	Backbone	mAP ↑	MFLOPs/ev ↓	Parameter
AEGNN+YOLOX+S	Event	GNN	0.339	3.53	33.9M
AEGNN+YOLOX+L	Event	GNN+SSM	0.362	3.67	37.7M
EHGCN-S	Event	GNN	0.394	2.98	14.1M
EHGCN-L	Event	GNN+SSM	0.427	3.10	17.9M
SMamba	Event	RNN+SSM	0.621	**0.48**	16.1M
DAGr-S+ResNet-50	RGB+Event	GNN+CNN	0.682	6.77	34.4M
DAGr-S+ResNet-50*	RGB+Event	GNN+CNN	0.663	0.59	34.4M
Ours	RGB+Event	RNN+SSM	**0.743**	0.55	10.8M

We compare the proposed method with representative event-based detectors and RGB-Event multimodal methods, as shown in Tab. 3. For AEGNN+YOLOX-S and AEGNN+YOLOX-L Schaefer et al. (2022), we follow the comparison settings of EHGCN Chen et al. (2025a). Overall, event-only methods show relatively limited performance on EVMars-Harsh, as static rocks generate fewer informative events during short stops or low-speed rover motion, while dust and illumination changes may continue to trigger noisy responses. SMamba achieves the best performance among event-only methods by combining temporal information propagation with ConvLSTM and global dependency modeling based on Mamba/SSM, but it remains limited by the lack of stable appearance cues under sparse or noisy event responses. In contrast, our method jointly exploits RGB and event information to achieve 0.743 mAP and outperforms the RGB-Event DAGr variants Gehrig & Scaramuzza (2024). Meanwhile, it uses only 10.8 M parameters and 0.55 MFLOPs/ev, demonstrating a favorable balance between detection accuracy and computational complexity.

6 Conclusion

This work addresses robust object detection in extreme environments by constructing **EVMars-Harsh**, an RGB-Event multimodal dataset in Mars-like dusty scenes, and proposing a multimodal detection framework tailored to perception degradation. The proposed method employs DAS to suppress dust-induced random events, RAMF to adaptively fuse RGB, Color Event, and Polarity Event information, and RA-LASSM to perform reliability-guided long-range feature modeling. Experimental results show that the proposed method achieves **0.743 mAP** on EVMars-Harsh while maintaining relatively low model complexity and event-level computational cost. Future work will further investigate its generalization to broader extreme scenarios and its deployment in resource-constrained real-time systems.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, sentence refinement, and improving the readability and conciseness of parts of the manuscript. Generative AI was not used to generate experimental results, formulate mathematical claims, design the core methodology, or interpret the experimental findings. All AI-assisted text was carefully reviewed and revised by the authors to ensure technical accuracy, consistency with the conducted research, and appropriate attribution to the relevant literature. The authors take full responsibility for the final content of this work, including all text, claims, experimental results, and conclusions.

REFERENCES

Haosheng Chen, Lian Luo, Mengjingcheng Mo, Zhanjie Wu, Ji Gan, Jiaxu Leng, and Xinbo Gao. Ehgcn: Hierarchical euclidean-hyperbolic fusion via motion-aware gcn for hybrid event stream perception. *arXiv preprint arXiv:2504.16616*, 2025a.

Nuo Chen, Chao Xiao, Yimian Dai, Shiman He, Miao Li, and Wei An. Event-based tiny object detection: A benchmark dataset and baseline. In *the IEEE International Conference on Computer Vision (ICCV)*, pp. 7209–7218, 2025b.

Pierre de Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. arXiv preprint, arXiv:2001.08499, 2020.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024.

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.

Haiqing Hao, Nikola Zubic, Weihua He, Zhipeng Sui, Davide Scaramuzza, and Wenhui Wang. Maximizing asynchronicity in event-based neural networks. In *International Conference on Learning Representations*, volume 2026, pp. 123834–123855, 2026.

Elisabeth M Hausrath, Alexis K Navarre-Sitchler, Peter B Sak, Carl I Steefel, and Susan L Brantley. Basalt weathering rates on earth and the duration of liquid water on the plains of gusev crater, mars. *Geology*, 36(1):67–70, 2008.

Dianze Li, Yonghong Tian, and Jianing Li. Sodformer: Streaming object detection with transformer using events and frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (11):14020–14037, 2023.

Zhanwen Liu, Nan Yang, Yang Wang, Yuke Li, Xiangmo Zhao, and Fei-Yue Wang. Enhancing traffic object detection in variable illumination with rgb-event fusion. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):20335–20350, 2024.

Yunfan Lu, Yanlin Qian, Ziyang Rao, Junren Xiao, Liming Chen, and Hui Xiong. Rgb-event isp: The dataset and benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 16556–16570, 2025.

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 *International Conference on Learning Representations*, 2026.

Florian Mahlknecht, Daniel Gehrig, Jeremy Nash, Friedrich M Rockenbauer, Benjamin Morrell, Jeff Delaune, and Davide Scaramuzza. Exploring event camera-based odometry for planetary robots. *IEEE Robotics and Automation Letters*, 7(4):8651–8658, 2022.

Harry Y McSween, Raymond E Arvidson, James F Bell III, D Blaney, Nathalie A Cabrol, Philip R Christensen, Benton C Clark, JA Crisp, Larry S Crumpler, DJ Des Marais, et al. Basaltic rocks analyzed by the spirit rover in gusev crater. *Science*, 305(5685):842–845, 2004.

Harry Y McSween Jr. Petrology on mars. *American Mineralogist*, 100(11-12):2380–2395, 2015.

Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. In *the Neural Information Processing Systems*, pp. 16639–16652, 2020.

Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. Aegnn: Asynchronous event-based graph neural networks. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12371–12381, 2022.

Mingyuan Sun, Donghao Zhang, Zongyuan Ge, Jiaxu Wang, Jia Li, Zheng Fang, and Renjing Xu. Eventrpg: Event data augmentation with relevance propagation guidance. In *International Conference on Learning Representations*, volume 2024, pp. 6366–6385, 2024.

David T Vaniman, DL Bish, DW Ming, TF Bristow, RV Morris, DF Blake, SJ Chipera, SM Morrison, AH Treiman, EB Rampe, et al. Mineralogy of a mudstone at yellowknife bay, gale crater, mars. *Science*, 343(6169):1243480, 2014.

Aayush Atul Verma, Bharatesh Chakravarthi, Arpitsinh Vaghela, Hua Wei, and Yezhou Yang. etram: Event-based traffic monitoring dataset. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 22637–22646, 2024.

Xiao Wang, Yu Jin, Wentao Wu, Wei Zhang, Lin Zhu, Bo Jiang, and Yonghong Tian. Object detection using event camera: A moe heat conduction based detector and a new benchmark dataset. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 29321–29330, 2025.

Michael B Wyatt and Harry Y McSween Jr. Spectral evidence for weathered basalt as an alternative to andesite in the northern lowlands of mars. *Nature*, 417(6886):263–266, 2002.

AS Yen, DW Ming, DT Vaniman, R Gellert, DF Blake, RV Morris, SM Morrison, TF Bristow, SJ Chipera, KS Edgett, et al. Multiple stages of aqueous alteration along fractures in mudstone and sandstone strata in gale crater, mars. *Earth and Planetary Science Letters*, 471:186–198, 2017.

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 *International Conference on Learning Representations*, 2026.

Hanyu Zhou, Yi Chang, Haoyue Liu, YAN WENDING, Yuxing Duan, Zhiwei Shi, and Luxin Yan. Exploring the common appearance-boundary adaptation for nighttime optical flow. In *International Conference on Learning Representations*, volume 2024, pp. 33145–33155, 2024.

Zhuyun Zhou, Zongwei Wu, Rémi Boutteau, Fan Yang, Cédric Demonceaux, and Dominique Ginhac. Rgb-event fusion for moving object detection in autonomous driving. In *the IEEE International Conference on Robotics and Automation*, pp. 7808–7815, 2023.