SACFUSION: FROM SEMANTIC COMPATIBILITY TO CONDITIONAL COMPLEMENTARITY IN RGB-EVENT OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB-Event object detection leverages complementary information to improve robust perception under challenging and highly dynamic conditions, particularly for autonomous driving. However, substantial differences between RGB and Event modalities in semantic representation, observation density, and information reliability often lead to modality bias during fusion. To address this issue, we propose SACFusion, a two stage RGB-Event object detection framework that progresses from semantic compatibility to conditional complementarity. In Stage I, a Temporal Polarity Aggregator explicitly models the temporal, polarity, and local spatial structures of Event data, while Scale Aware Local Semantic Alignment and Cross Path Semantic Preservation adapt Event features to the pretrained RGB semantic space without reconstructing RGB appearance. In Stage II, an RGB dominant asymmetric fusion mechanism uses a spatially varying Event Retrieval Prior to guide RGB queries toward more informative Event regions at corresponding feature scales. This design preserves stable RGB representations while reducing interference from sparse or unreliable Event observations. By separating representation adaptation from complementary interaction, SACFusion enables Event information to be exploited more selectively according to local scene conditions. Experiments on DSEC-Det and PEOD demonstrate consistent improvements under diverse and challenging conditions, achieving relative mAP gains of 11.1% and 16.5%, respectively, over the strongest competing methods.

1 INTRODUCTION

Event cameras offer high temporal resolution, a wide dynamic range, and robustness to motion blur, making them well suited to challenging visual conditions (Gallego et al., 2020; Duan et al., 2025; Zhang et al., 2025). In relatively constrained applications such as UAV detection, target motion can induce distinctive spatiotemporal Event patterns, providing motion-sensitive cues that complement RGB appearance (Chen et al., 2025; He et al., 2026). Autonomous driving is substantially more complex, where ego-motion, dynamic traffic, depth variation, and changing illumination generate widespread and heterogeneous Event responses (Gehrig et al., 2021). Consequently, the usefulness of Event information varies substantially across spatial locations and scene conditions (Kong et al., 2025). Existing methods commonly employ feature concatenation, attention, or Transformer based interaction, but directly applying these mechanisms to heterogeneous RGB and Event representations may be suboptimal when their representation quality and local reliability differ significantly (Wang et al., 2026b; Li et al., 2023).

The difficulty therefore lies not in the interaction mechanisms themselves, but in the unequal representations on which they operate (Huang et al., 2022; Wang et al., 2026a). RGB branches benefit from mature representations learned from large scale image datasets, whereas Event branches rely on substantially less training data and weaker pretrained priors, resulting in a clear representation imbalance (Yang et al., 2023; 2025). This imbalance can be further amplified during joint optimization, as illustrated in Fig. 1, where the stronger RGB branch tends to dominate the shared objective and leaves the Event representation underoptimized. When such immature Event features are directly involved in symmetric fusion, they may interfere with well established RGB representations and cause negative transfer (Peng et al., 2022). We refer to these coupled effects as modality bias,

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

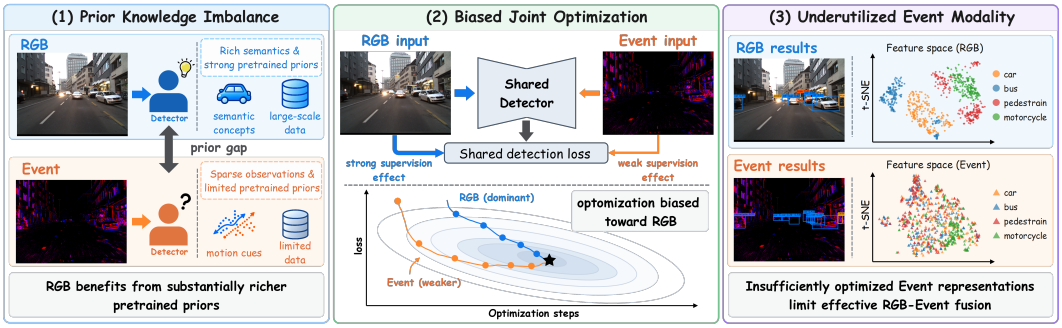

Figure 1: Illustration of modality bias in RGB-Event object detection. Richer RGB pretrained priors lead to imbalanced optimization and underoptimized Event representations, ultimately limiting effective multimodal fusion. The t-SNE visualization is based on features extracted from the DSEC-Det dataset.

consisting of representation bias from unequal semantic capacities and interaction bias from indiscriminate multimodal interaction.

Recent studies have begun to address the challenges caused by the discrepancy between RGB and Event representations from different perspectives (Sun et al., 2024). SFNet (Liu et al., 2024) improves Event representations by transforming sparse events into more structured visual features, while MCFNet (Liu et al., 2025) enhances temporal and spatial correspondence through optical flow based correction. These methods primarily aim to transform Event information into representations that are closer to RGB semantics before being fed into the subsequent fusion detector, whereas SPFD (Wang et al., 2026b) focuses more on cross modal interaction through shared and modality specific feature decomposition. Together, these studies suggest that effective RGB-Event detection requires both representation compatibility and appropriate cross modal interaction.

Based on this observation, we further decompose modality bias in RGB-Event object detection into two related problems: representation bias and interaction bias. Representation bias arises from the substantial differences in sensing mechanisms and feature distributions between RGB and Event data, which make unconstrained Event features difficult to adapt directly to the detection semantics learned from large RGB datasets. Interaction bias, in contrast, stems from nearly equal or indiscriminate information exchange during fusion, allowing an unstable auxiliary modality to continuously affect the main RGB representation. We therefore argue that semantic compatibility and conditional complementarity should be treated as two distinct but progressive learning objectives. Event representations should first become compatible with the semantic space of the RGB detector, after which the model should determine which Event evidence provides meaningful complementary information for the current RGB representation.

Based on this perspective, we propose SACFusion, a two stage RGB-Event object detection framework built on DINO to progressively address representation bias and interaction bias. Rather than directly optimizing heterogeneous modalities through symmetric fusion, SACFusion first establishes semantic compatibility and then learns conditional complementarity. In Stage I, the Event branch is adapted toward the pretrained RGB semantic space through Temporal Polarity Aggregation (TPA), Scale Aware Local Semantic Alignment (SLSA), and Cross Path Semantic Preservation (CP), reducing the representation imbalance before fusion. Stage II then adopts an RGB dominant asymmetric interaction scheme, where RGB features serve as queries and Event features provide auxiliary evidence. A learned Event Retrieval Prior further guides local retrieval toward more informative Event regions. In this way, SACFusion separates representation adaptation from complementary interaction, allowing the two sources of modality bias to be addressed progressively.

The main contributions of this work are summarized as follows:

- Two coupled forms of modality bias in RGB-Event detection are identified, motivating semantic compatibility before conditional complementarity.
- A semantic compatibility learning scheme integrates TPA, SLSA, and CP to progressively adapt Event representations to the RGB semantic space.

- An RGB dominant asymmetric fusion mechanism uses a spatially varying Event Retrieval Prior to selectively retrieve complementary Event evidence while reducing unreliable cross modal interference.

2 Related Work

2.1 RGB-Event Object Detection

RGB-Event object detection combines the texture and semantic information of RGB images with the dynamic response of Event cameras under rapid motion and extreme illumination (Gehrig & Scaramuzza, 2024), improving perception robustness in challenging scenes. Early studies mainly focused on multi scale feature fusion and temporal modeling of events. For example, FPN-Fusion (Tomy et al., 2022) employs a dual branch architecture to fuse RGB and Event features at multiple levels of the feature pyramid, while RENet (Zhou et al., 2023) further enhances information exchange between the two modalities through temporal multi scale event representations and bidirectional interaction. With the development of Transformers, SODFormer (Li et al., 2023) enables continuous RGB-Event detection through spatiotemporal modeling and asynchronous attention, while EOLO (Cao et al., 2024b) combines temporal event modeling with adaptive fusion to improve detection under diverse illumination conditions.

Recent studies have gradually shifted from simply increasing cross modal interaction toward improving Event representation quality and explicitly modeling modality differences. SFNet (Liu et al., 2024) constructs more structurally complete event representations to facilitate subsequent fusion. CAFR (Cao et al., 2024a) enhances modality collaboration through coarse to fine feature refinement and statistical alignment, while MCFNet (Liu et al., 2025) further improves Event representations with motion guided temporal correction and spatial enhancement. SPFD (Wang et al., 2026b) explicitly separates shared and modality specific features, selectively exploiting RGB texture and Event motion cues across different regions. These studies indicate that effective RGB-Event detection depends on both representation compatibility and cross modal interaction. Unlike direct joint fusion, our method first establishes semantic compatibility and then selectively retrieves complementary Event evidence to refine RGB features.

2.2 Modality Imbalance in Multimodal Learning

Different sensing modalities often differ substantially in imaging mechanisms, information density, and task capability, making fully symmetric interaction a poor fit for heterogeneous multimodal fusion. In RGB LiDAR 3D detection, LiDAR provides reliable geometry and depth, while RGB offers richer texture and semantic cues (Wang et al., 2024). Recent methods such as TransFusion (Bai et al., 2022) and GAFusion (Li et al., 2024) therefore adopt query based or selective aggregation strategies, using stable object or BEV representations to retrieve complementary image information when needed rather than mixing features indiscriminately. A similar trend appears in RGB Thermal and infrared perception, where modality reliability varies with illumination and environmental conditions. Dynamic weighting, quality aware fusion, and primary auxiliary modality designs have been explored to reduce interference from unreliable inputs (Wang et al., 2026c; Zhu et al., 2025).

More broadly, this issue is often described as modality imbalance in multimodal learning, where modalities with stronger discriminative capability or faster convergence can dominate joint optimization and leave weaker modalities insufficiently learned (Huang et al., 2022; Zong et al., 2024). Existing studies have therefore explored gradient modulation, learning-rate rebalancing, and modality-specific regularization to mitigate such optimization imbalance (Peng et al., 2022; Fan et al., 2023). These methods, however, primarily address optimization imbalance among already established modality representations or variations in their relative learning dynamics. RGB-Event detection introduces an additional representation challenge because the two modalities differ fundamentally in signal formation and semantic maturity; in particular, large-scale and transferable pretraining is far less established for Event data than for conventional RGB imagery. (Yang et al., 2023). Recent RGB-Event studies further suggest that the two modalities contribute differently across spatial regions and semantic contexts (Wang et al., 2026b). Motivated by these observations, we separately address representation and interaction biases, progressing from semantic compatibility to conditional complementarity.

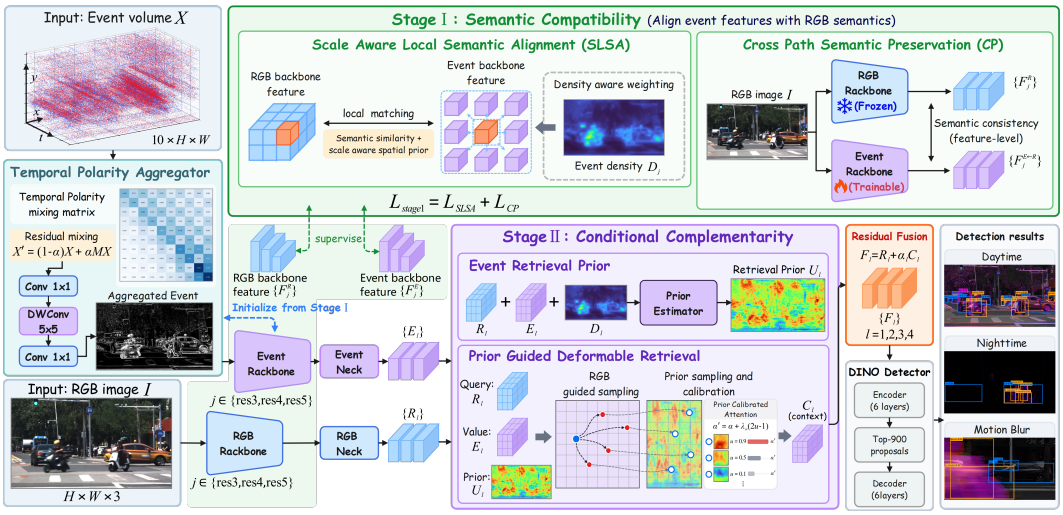

Figure 2: Overview of SACFusion. Stage I establishes semantic compatibility between RGB and Event representations, while Stage II learns conditional complementarity through Event retrieval prior guided asymmetric fusion.

3 METHOD

Given a synchronized RGB image I and an event voxel X, SACFusion aims to exploit complementary Event cues under fast motion and visual degradation without disturbing stable RGB representations. It follows a two stage design that progresses from semantic compatibility to conditional complementarity.

3.1 OVERALL FRAMEWORK

As illustrated in Fig. 2, SACFusion consists of temporal and polarity event aggregation, cross modal semantic alignment, asymmetric feature fusion, and a DINO detector. Given an RGB image I and an event voxel $X \in \mathbb{R}^{B\times 2T\times H\times W}$ with $T = 5$ temporal bins and two polarities, the RGB backbone $\mathcal{B}_R$ directly extracts visual features, while the Event input is first converted into a three channel representation by the temporal and polarity aggregator $\mathcal{A}$. The modality specific backbones produce $\{F_j^R\}_{j\in\mathcal{S}} = \mathcal{B}_R(I)$ and $\{F_j^E\}_{j\in\mathcal{S}} = \mathcal{B}_E(\mathcal{A}(X))$, where $\mathcal{S} = \{3, 4, 5\}$ denotes the backbone levels.

SACFusion is trained in two stages. Stage I freezes the RGB backbone and aligns the Event branch with its pretrained semantic space. Stage II initializes from the aligned dual branch network and employs independent necks to construct four scale RGB and Event features $\{R_l, E_l\}_{l=1}^4$. At each scale, RGB features serve as queries to retrieve complementary Event context C_l under the guidance of a learned Event retrieval prior. The fused representation is

$$F_l = R_l + \alpha_l C_l, \tag{1}$$

where α_l is a learnable residual coefficient. The resulting features $\{F_l\}_{l=1}^4$ are finally fed into the standard DINO encoder and decoder for object detection.

3.2 SEMANTIC COMPATIBILITY LEARNING

Stage I aims to adapt Event representations to the pretrained RGB semantic space before multimodal fusion. The RGB backbone is frozen and serves as a semantic reference, while the Event branch is optimized through temporal polarity aggregation, local semantic alignment, and cross path semantic preservation.

3.2.1 TEMPORAL POLARITY EVENT AGGREGATION

The event voxel jointly encodes temporal and polarity structures, which may be weakened by directly compressing the $2T$ channels into an image representation. We therefore introduce a

lightweight Temporal Polarity Aggregator (TPA) that explicitly propagates information across temporal bins and polarities before spatial aggregation.

For event channels c and d with temporal indices and polarities (t_c, p_c) and (t_d, p_d), respectively, we define the temporal polarity association weight as

$$\widetilde{M}_{cd} = \beta_{p_c,p_d} \exp\left(-\frac{|t_c - t_d|}{\sigma_{p_c,p_d}}\right) \exp\left(-\frac{T-1-t_d}{\tau}\right), \qquad M_{cd} = \frac{\widetilde{M}_{cd}}{\sum_{d'} \widetilde{M}_{cd'} + \epsilon}. \tag{2}$$

Here, σ_{p_c,p_d} and β_{p_c,p_d} control temporal decay and polarity interaction strength, while τ favors more recent events.

The normalized associations are used for channel propagation and residual mixing,

$$\widetilde{X}_c = \sum_d M_{cd} X_d, \qquad X'_c = (1-\gamma) X_c + \gamma \widetilde{X}_c. \tag{3}$$

The mixed features are then processed by a 1×1 projection, a 5×5 depthwise convolution, and a final 1×1 projection. The resulting three channel representation $X_{\text{agg}} = \mathcal{A}(X)$ preserves temporal, polarity, and local spatial structures while matching the input format of the pretrained Event backbone.

3.2.2 Scale Aware Local Semantic Alignment

To accommodate slight spatial misalignment between RGB and Event observations, we introduce Scale Aware Local Semantic Alignment (SLSA). For each RGB feature at location x and backbone level $j \in \mathcal{S} = \{3, 4, 5\}$, SLSA searches for semantic correspondence within a 3×3 Event neighborhood Δ. Using channel normalized features $\bar{F}_j^R$ and $\bar{F}_j^E$, the matching score is

$$q_j(x, \delta) = \frac{\bar{F}_j^R(x)^\top \bar{F}_j^E(x+\delta)}{\tau_s} - \frac{\|s_j \delta\|_2^2}{2\sigma_{\text{sp}}^2}, \tag{4}$$

where $\delta \in \Delta$, τ_s is the matching temperature, s_j denotes the feature stride, and σ_{sp} controls spatial tolerance. The locally aligned Event feature is

$$\widehat{F}_j^E(x) = \sum_{\delta \in \Delta} a_j(x, \delta) F_j^E(x+\delta), \qquad a_j(x, \delta) = \text{softmax}_\delta\big(q_j(x, \delta)\big). \tag{5}$$

To emphasize informative Event observations, the resized density map D_j is converted into a bounded density response $\rho_j(x) = 1 - \exp(-D_j(x)/\kappa)$. The alignment objective is

$$\mathcal{L}_{\text{SLSA}} = \sum_{j \in \mathcal{S}} \lambda_j \frac{\sum_x m_j(x) \rho_j(x) \left[1 - \cos\left(\widehat{F}_j^E(x), F_j^R(x)\right)\right]}{\max\left(\sum_x m_j(x) \rho_j(x), 1\right)}. \tag{6}$$

Here, $m_j(x)$ denotes the valid region mask excluding padded locations, and λ_j is the level wise alignment weight. The batch dimension is omitted for clarity.

3.2.3 Cross Path Semantic Preservation

Local alignment alone may cause the Event backbone to drift from the pretrained RGB semantic space. Following Cao et al. (2026), we introduce Cross Path Semantic Preservation (CP) to mitigate this drift. The RGB image is additionally passed through the Event backbone as $F_j^{E \leftarrow R} = [\mathcal{B}_E(I)]_j$ and constrained to match the frozen RGB representation,

$$\mathcal{L}_{\text{CP}} = \sum_{j \in \mathcal{S}} \frac{1}{|\Omega_j|} \sum_{x \in \Omega_j} \left[1 - \cos\left(F_j^{E \leftarrow R}(x), \text{sg}(F_j^R(x))\right)\right], \tag{7}$$

where $\text{sg}(\cdot)$ denotes stop gradient. The Stage I objective is $\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{SLSA}} + \mathcal{L}_{\text{CP}}$.

3.3 Asymmetric Conditional Complementarity Fusion

Semantic compatibility does not imply that Event features are equally informative across spatial locations. Stage II therefore learns conditional complementarity through RGB dominant asymmetric retrieval, where RGB features act as queries and Event features provide auxiliary evidence.

Table 1: Quantitative comparison on DSEC-Det and PEOD. The best results are highlighted in bold.

Method	Pub. & Year	DSEC-Det	PEOD			Param. (M)
			Normal	Challenge	Overall	
		$\mathrm{mAP}_{50}/\mathrm{mAP}$	$\mathrm{mAP}_{50}/\mathrm{mAP}$	$\mathrm{mAP}_{50}/\mathrm{mAP}$	$\mathrm{mAP}_{50}/\mathrm{mAP}$	
YOLOX-M	arXiv'21	51.8/30.8	62.7/38.4	38.6/19.5	47.5/26.7	25.3
YOLO2026-M	arXiv'26	45.8/28.2	58.4/38.1	32.2/18.4	42.1/25.7	20.4
DINO-R50	ICLR'23	53.8/32.3	64.5/42.2	48.8/25.9	54.5/32.1	47.0
SAST	CVPR'24	24.3/12.1	53.9/30.4	44.4/21.3	47.9/25.0	24.3
MvHeat-DET	CVPR'25	29.4/14.4	65.5/39.1	35.5/18.7	47.3/26.5	65.1
FPN-Fusion	ICRA'22	35.4/18.4	45.2/26.0	17.0/8.4	27.9/15.0	65.6
EOLO	ICRA'24	33.9/19.6	51.5/31.9	20.6/10.1	32.7/18.6	46.3
SFNet	T-ITS'24	48.0/27.9	56.9/32.7	36.6/18.3	42.9/23.0	57.5
CAFR	ECCV'24	34.8/17.8	60.3/35.9	31.6/14.9	43.1/23.4	82.5
SPFD	CVPR'26	54.1/32.2	62.8/39.6	50.0/27.8	52.2/31.3	78.6
SACFusion (Ours)	–	**59.3/35.9**	**70.3/46.0**	**55.5/31.8**	**60.9/37.4**	77.4

3.3.1 Event Retrieval Prior Estimation

At each fusion scale l, RGB and Event features are projected into a shared 64 dimensional relation space,

$$r_l = \mathbf{W}_s^{(l)}\,\mathrm{GN}(R_l), \qquad e_l = \mathbf{W}_s^{(l)}\,\mathrm{GN}(E_l), \tag{8}$$

where $\mathbf{W}_s^{(l)}$ is shared between the two modalities at the same scale. The resized Event density D_l is transformed using the same bounded mapping, $\rho_l = 1 - \exp(-D_l/\kappa)$. We then construct

$$Z_l = [r_l,\, e_l,\, e_l - r_l,\, |e_l - r_l|,\, r_l \odot e_l,\, \rho_l]\,, \qquad U_l = \mathcal{G}_l(Z_l). \tag{9}$$

The scale specific estimator $\mathcal{G}_l$ captures cross modal semantic relations and Event observation strength to provide spatial guidance for subsequent retrieval.

3.3.2 Prior Guided Event Retrieval

Cross modal interaction is restricted to corresponding scales, with R_l querying E_l. For each RGB query position i, attention head h predicts K offsets and samples Event features and retrieval priors at the same locations,

$$\begin{aligned} &\mathbf{p}_{i,h,k} = \mathbf{p}_i^{\mathrm{ref}} + \Delta\mathbf{p}_{i,h,k} \oslash [W_l, H_l]^\top, \\ &\mathbf{v}_{i,h,k} = \mathrm{Sample}(\mathbf{W}_v E_l, \mathbf{p}_{i,h,k}), \qquad u_{i,h,k} = \mathrm{Sample}(U_l, \mathbf{p}_{i,h,k}). \end{aligned} \tag{10}$$

The sampled prior calibrates the RGB predicted attention logits,

$$a'_{i,h,k} = a_{i,h,k} + \lambda_u(2u_{i,h,k} - 1), \qquad w_{i,h,k} = \mathrm{softmax}_k(a'_{i,h,k}), \tag{11}$$

thereby prioritizing Event candidates with higher retrieval relevance. The resulting context is aggregated as

$$C_{l,i} = \mathbf{W}_o\,\mathrm{Concat}_h\left(\sum_{k=1}^{K} w_{i,h,k}\mathbf{v}_{i,h,k}\right), \tag{12}$$

and incorporated into the RGB representation through Eq. 1.

The interaction is asymmetric by design. RGB features determine the queries and sampling locations, while Event features provide auxiliary evidence for residual refinement. The Event Retrieval Prior U_l reweights sampled candidates rather than controlling global Event injection.

3.4 Training Strategy

SACFusion is optimized in two stages. Stage I freezes the RGB backbone and optimizes the Event aggregator and Event backbone using $\mathcal{L}_{\mathrm{SLSA}} + \mathcal{L}_{\mathrm{CP}}$. Stage II initializes from the aligned dual branch network, disables SLSA and CP, and jointly optimizes the full detector using the standard DINO detection objective. The Event Retrieval Prior receives no explicit supervision and is learned end to end from the detection loss.

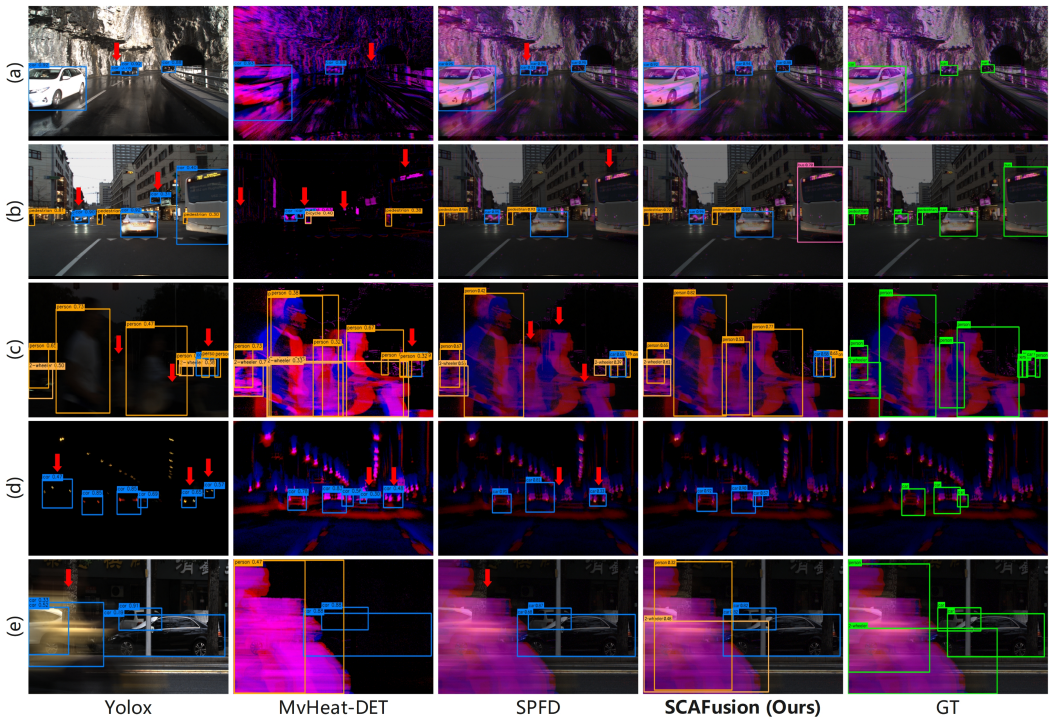

Figure 3: Qualitative comparison on DSEC-Det and PEOD. From left to right, the results of YOLOX-M, MvHeat-DET, SPFD, Ours, and ground truth (GT) are shown. Red arrows highlight representative missed or inaccurate detections.

Table 2: Ablation study of SACFusion components on DSEC-Det. Naive Fusion averages the RGB and Event neck features with equal weights of 0.5.

Setting	TPA	SLSA + CP	Asymmetric Retrieval	AP	AP_{50}	AP_{75}
RGB-only	–	–	–	31.84	53.15	32.55
Naive Fusion	–	–	–	33.21	57.66	32.63
+ TPA	✓	–	–	34.21	56.70	35.19
+ Semantic Alignment	✓	✓	–	35.16	59.14	35.77
SACFusion (Full)	✓	✓	✓	**35.89**	**59.32**	**36.87**

4 Experiments

4.1 Experimental Settings

Datasets. We conduct experiments on two public RGB-Event object detection datasets in autonomous driving scenarios, DSEC-Det (Liu et al., 2024) and PEOD (Cui et al., 2026). DSEC-Det contains 60 driving sequences, 70,379 annotated frames, and approximately 390k bounding boxes across 8 object categories, with an Event resolution of 640×480. PEOD provides approximately 72k pixel aligned RGB-Event pairs and 340k bounding boxes covering 6 traffic object categories, with an original resolution of 1280×720. A substantial portion of PEOD is collected under challenging conditions such as low illumination, overexposure, and high speed motion.

Baselines. We compare SACFusion with representative RGB, Event, and RGB-Event detection methods. The RGB baselines include YOLOX (Ge et al., 2021), YOLO26 (Jocher et al., 2026), and DINO (Zhang et al., 2023). Event based detectors include SAST (Peng et al., 2024) and MvHeat-DET (Wang et al., 2025), while RGB-Event fusion methods include FPN-Fusion (Tomy et al., 2022), EOLO (Cao et al., 2024b), SFNet (Liu et al., 2024), CAFR (Cao et al., 2024a), and SPFD (Wang et al., 2026b). These methods cover convolutional architectures, Transformer based detectors, and different cross modal fusion strategies.

Evaluation Metrics. We follow the standard COCO object detection evaluation protocol and report mAP, $\mathrm{mAP50}$, and $\mathrm{mAP75}$. Here, mAP denotes the mean AP averaged over IoU thresholds from

0.50 to 0.95 with a step size of 0.05 and is used as the primary evaluation metric. $\mathrm{mAP50}$ and $\mathrm{mAP75}$ denote the detection performance at IoU thresholds of 0.50 and 0.75, respectively.

Implementation Details. SACFusion is trained on 8 NVIDIA RTX 4090 GPUs. Both RGB and Event branches use ResNet-50 backbones initialized from a COCO pretrained DINO-R50 model. In Stage I, the RGB backbone is frozen, while the Event branch and TPA are optimized with SLSA and CP for 30k iterations using a batch size of 40 and a learning rate of 1×10^{-4}. In Stage II, the aligned weights are used to initialize end to end detection training for 20k iterations with a batch size of 24. The pretrained backbones and TPA use a learning rate of 1×10^{-5}, while the remaining modules use 1×10^{-4}. AdamW (Loshchilov & Hutter, 2019) with cosine learning rate decay is used in both stages.

4.2 Comparison with SOTA methods

Quantitative Comparison. SACFusion consistently achieves the strongest performance across both benchmarks, as summarized in Table 1. On DSEC-Det, it reaches $35.9\,\mathrm{mAP}$, exceeding the best competing method by 3.6 percentage points. On PEOD, SACFusion obtains 46.0, 31.8, and $37.4\,\mathrm{mAP}$ under the Normal, Challenge, and Overall settings, respectively, corresponding to gains of 3.8, 4.0, and 5.3 points over the strongest competing results. The improvement is particularly evident under challenging conditions, where RGB appearance becomes less reliable and complementary Event information is more valuable. These consistent gains across datasets and operating conditions indicate that SACFusion can better exploit Event evidence without sacrificing the stability of RGB representations. Meanwhile, its parameter count remains comparable to existing RGB-Event fusion methods, showing that the performance gain does not rely on a substantially larger model.

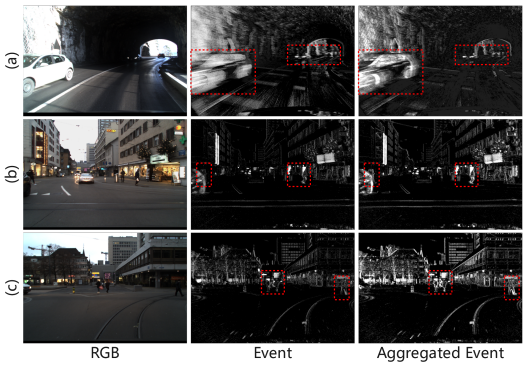

Figure 4: Visualization of the proposed TPA. From left to right, RGB images, raw Event representations, and aggregated Event representations are shown.

Qualitative Comparison. SACFusion exhibits more robust detection across diverse challenging scenes, with representative examples presented in Fig. 3. In the tunnel and urban scenes, YOLOX-M misses several distant or weakly visible vehicles and pedestrians despite retaining reliable detections for visually salient objects. Under severe low illumination and motion degradation, this limitation becomes more pronounced, with multiple targets indicated by the red arrows being missed or inaccurately localized. MvHeat-DET preserves stronger responses to moving objects and can recover some targets that are difficult to recognize from RGB appearance alone, but its predictions are less stable in terms of localization and category assignment. SPFD benefits from multimodal fusion and improves over the unimodal baselines, yet several difficult targets remain undetected. In comparison, SACFusion recovers more of the missed vehicles and traffic participants while maintaining localization that is more consistent with the ground truth. These results support the effectiveness of semantic compatibility learning and asymmetric Event retrieval in exploiting useful Event evidence when RGB appearance becomes unreliable.

4.3 Ablation Study

We progressively introduce the proposed components to examine their individual contributions, with the results summarized in Table 2. Naive RGB-Event fusion improves the RGB-only baseline from 31.84 to $33.21\,\mathrm{AP}$, confirming the complementary value of Event observations. Incorporating TPA further raises the performance to $34.21\,\mathrm{AP}$, showing the benefit of explicitly modeling temporal and polarity structures before feature extraction. Semantic Alignment brings an additional improvement to $35.16\,\mathrm{AP}$, supporting the importance of reducing the representation gap before multimodal interaction. Finally, the proposed Asymmetric Fusion achieves $35.89\,\mathrm{AP}$ and $36.87\,\mathrm{AP}_{75}$, yielding an overall gain of $4.05\,\mathrm{AP}$ over the RGB-only baseline. These consistent improvements indicate that event aggregation, semantic compatibility learning, and conditional complementarity contribute progressively to the final performance.

4.4 Effectiveness Analysis

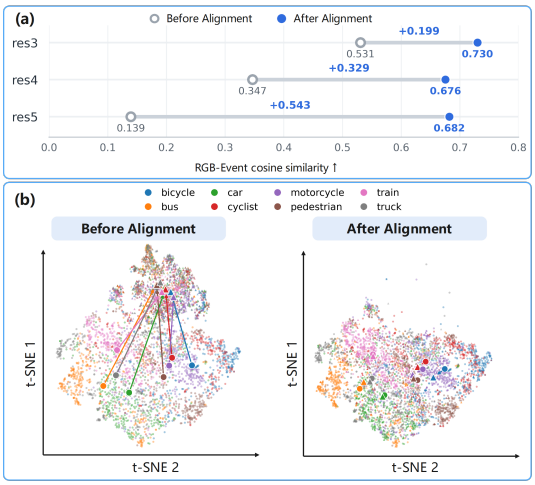

Figure 5: Effect of semantic alignment on RGB-Event representations. (a) Cosine similarity across backbone levels before and after alignment. (b) t-SNE visualization of the corresponding feature distributions.

Temporal Polarity Aggregation. The proposed TPA produces more structured Event representations by explicitly modeling temporal and polarity dependencies. Compared with raw Event inputs, the aggregated representations exhibit more concentrated responses around moving objects and clearer structural boundaries, while scattered event activations are noticeably reduced, as illustrated in Fig. 4. This suggests that TPA preserves informative motion structures while providing a more suitable representation for the subsequent Event backbone.

Semantic Compatibility. Stage I consistently improves the semantic agreement between RGB and Event representations. The cosine similarity increases from 0.531 to 0.730 at res3, from 0.347 to 0.676 at res4, and from 0.139 to 0.682 at res5. The improvement becomes more pronounced at deeper feature levels, indicating that the proposed alignment is particularly effective in reducing high level semantic discrepancies. The corresponding t-SNE visualization further shows a closer correspondence between RGB and Event feature distributions after alignment, as shown in Fig. 5. These observations support the role of SLSA and CP in establishing semantic compatibility before fusion.

Event Retrieval Prior. The learned Event Retrieval Prior exhibits clear spatial variation across different visual conditions. Regions containing informative Event structures tend to receive stronger responses, whereas less useful regions are assigned lower priority. This behavior remains evident under challenging cases such as fast motion, nighttime illumination, and severe overexposure, where the prior emphasizes vehicle contours and other structurally informative Event responses despite degraded RGB appearance. These spatially varying responses are sampled together with the Event features during deformable retrieval, allowing the prior to bias attention toward more informative candidate locations. The visualization in Fig. 6 further demonstrates that the model learns task relevant spatial guidance for selective Event retrieval.

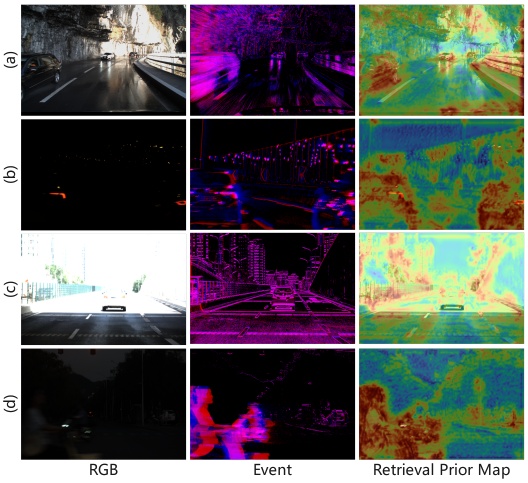

Figure 6: Visualization of the learned Event retrieval prior. From left to right, RGB images, Event representations, and the corresponding retrieval prior maps are shown.

5 Conclusions

SACFusion demonstrates that RGB-Event fusion benefits from explicitly separating representation adaptation from complementary interaction. Across DSEC-Det and PEOD, the proposed framework achieves consistent improvements over existing methods, including under challenging visual conditions. Ablation studies confirm the progressive contribution of temporal polarity aggregation, semantic alignment, and asymmetric fusion, while further analyses show improved RGB-Event feature consistency and meaningful spatial selectivity in the learned Event Retrieval Prior. These findings support the effectiveness of addressing modality bias through a compatibility first and complementarity aware design, providing a practical direction for robust RGB-Event perception.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and polishing of the manuscript to improve clarity and readability, as well as with software coding tasks such as code editing, debugging, and implementation assistance. Generative AI tools were not used to formulate the research hypotheses, design the proposed methodology or experiments, interpret experimental results, or draw scientific conclusions. All AI assisted text was reviewed and revised by the authors, and AI assisted code was manually inspected and tested to ensure correctness. We take full responsibility for the final content of this work, including all text, claims, experimental results, and software artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is conducted for academic research purposes and uses publicly available datasets. It does not involve human subject experiments, collection of private or sensitive information, or personally identifiable data. The proposed method is intended for research on RGB-Event perception, and any deployment in safety critical applications such as autonomous driving should undergo appropriate validation and evaluation. We are not aware of any additional ethical concerns associated with this work.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our results by providing detailed descriptions of the proposed method, model architecture, training settings, implementation details, datasets, and evaluation protocols in the main paper and appendix. The source code and relevant configuration files will be made publicly available upon acceptance.

REFERENCES

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1090–1099, 2022.

Hu Cao, Zehua Zhang, Yan Xia, Xinyi Li, Jiahao Xia, Guang Chen, and Alois Knoll. Embracing events and frames with hierarchical feature refinement network for object detection. In *European Conference on Computer Vision*, pp. 161–177. Springer, 2024a.

Jiahang Cao, Xu Zheng, Yuanhuiyi Lyu, Jiaxu Wang, Renjing Xu, and Lin Wang. Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9026–9032. IEEE, 2024b.

Jianwen Cao, Jiaxu Xing, Nico Messikommer, and Davide Scaramuzza. Generative event pretraining with foundation model alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 3189–3199, June 2026.

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

Luoping Cui, Hanqing Liu, Mingjie Liu, Endian Lin, Donghong Jiang, Yuhao Wang, and Chuang Zhu. Peod: A pixel-aligned event-rgb benchmark for object detection under challenging conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18207–18215, 2026.

Peiqi Duan, Boyu Li, Yixin Yang, Hanyue Lou, Minggui Teng, Xinyu Zhou, Yi Ma, and Boxin Shi. Eventaid: Benchmarking event-aided image/video enhancement algorithms with real-captured hybrid dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6959–6973, 2025.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20029–20038, June 2023.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 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.

Shiman He, Nuo Chen, Xinyi Ying, Yihang Luo, Yangsi Shi, Zaiping Lin, and Miao Li. Towards persistence: Learning topological constraints for event-based small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22291–22300, June 2026.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International conference on machine learning*, pp. 9226–9259. PMLR, 2022.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics yolo26: unified real-time end-to-end vision models. *arXiv preprint arXiv:2606.03748*, 2026.

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

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.

Xiaotian Li, Baojie Fan, Jiandong Tian, and Huijie Fan. Gafusion: Adaptive fusing lidar and camera with multiple guidance for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21209–21218, June 2024.

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.

Zhanwen Liu, Yujing Sun, Yang Wang, Nan Yang, Shengbo Eben Li, and Xiangmo Zhao. Beyond conventional vision: Rgb-event fusion for robust object detection in dynamic traffic scenarios. *Communications in Transportation Research*, 5:100202, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8237. IEEE, 2022.

Yansong Peng, Hebei Li, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. Scene adaptive sparse transformer for event-based object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16794–16804. IEEE, 2024.

Lei Sun, Daniel Gehrig, Christos Sakaridis, Mathias Gehrig, Jingyun Liang, Peng Sun, Zhijie Xu, Kaiwei Wang, Luc Van Gool, and Davide Scaramuzza. A unified framework for event-based frame interpolation with ad-hoc deblurring in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2265–2279, 2024.

Abhishek Tomy, Anshul Paigwar, Khushdeep S Mann, Alessandro Renzaglia, and Christian Laugier. Fusing event-based and rgb camera for robust object detection in adverse conditions. In *2022 International conference on robotics and automation (ICRA)*, pp. 933–939. IEEE, 2022.

Jian Wang, Fan Li, Yi An, Xuchong Zhang, and Hongbin Sun. Toward robust lidar-camera fusion in bev space via mutual deformable attention and temporal aggregation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):5753–5764, 2024.

Jun Wang, Fuyuan Cao, Zhixin Xue, Xingwang Zhao, and Jiye Liang. Cmob: Modality valuation via causal effect for balanced multimodal learning. *Advances in Neural Information Processing Systems*, 38:82585–82607, 2026a.

Keyao Wang, Shuai Liu, Hengda Shi, Lukui Shi, and Haiyong Chen. Beyond duality: A hybrid framework of leveraging shared and private features for rgb-event object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4415–4424, June 2026b.

Mianzhao Wang, Fan Shi, Xu Cheng, Chen Jia, and Shengyong Chen. Uncertainty-aware modality fusion for unaligned rgb-t salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41489–41498, June 2026c.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29321–29330. IEEE, 2025.

Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10665–10675. IEEE, 2023.

Yan Yang, Liyuan Pan, Dongxu Li, and Liu Liu. Ezsr: Event-based zero-shot recognition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4628–4638. IEEE, 2025.

Haitian Zhang, Chang Xu, Xinya Wang, Bingde Liu, Guang Hua, Lei Yu, and Wen Yang. Detecting every object from events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (8):7171–7178, 2025.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Haodong Zhu, Wenhao Dong, Linlin Yang, Hong Li, Yuguang Yang, Yangyang Ren, Qingcheng Zhu, Zichao Feng, Changbai Li, Shaohui Lin, Runqi Wang, Xiaoyan Luo, and Baochang Zhang. Wavemamba: Wavelet-driven mamba fusion for rgb-infrared object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11219–11229, October 2025.

Yongshuo Zong, Oisin Mac Aodha, and Timothy M Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5299–5318, 2024.

A Additional Implementation Details

A.1 Semantic Compatibility Learning

SLSA is applied to res3, res4, and res5 with spatial strides of 8, 16, and 32, respectively, using a fixed 3×3 local correspondence window $\Delta = \{-1, 0, 1\} \times \{-1, 0, 1\}$. The density sensitivity parameter is set to $\kappa = 1$ in all experiments. The valid region mask m_j is obtained by resizing the input validity mask to each feature resolution using nearest neighbor interpolation, so that padded locations do not contribute to the alignment loss. For CP, the RGB image is fed directly into the Event backbone without passing through TPA, ensuring that the constraint acts specifically on the semantic space of the Event backbone.

A.2 Conditional Complementarity Fusion

The four fusion scales contain 256 channels. For Event Retrieval Prior estimation, RGB and Event features are projected from 256 to 64 channels, with a separate estimator $\mathcal{G}_l$ used at each scale. Prior Guided Event Retrieval employs 8 attention heads and $K = 4$ sampling locations per head. Sampling offsets and base attention logits are predicted from RGB queries, while Event features and prior values are sampled at the same locations. We set $\lambda_u = 1$ in all experiments, and cross modal retrieval is restricted to corresponding feature scales.

A.3 Baseline Reproduction Details

To ensure a fair comparison, RGB-only and RGB-Event methods use pretrained weights only for their corresponding backbone components whenever applicable, while newly introduced necks, fusion modules, encoder decoder components, and detection heads are randomly initialized. This avoids giving RGB-only detectors an additional advantage from fully pretrained detection weights and reduces the influence of unequal pretraining data. Event-only methods are trained according to their original formulations.

All compared methods are reproduced as closely as possible following the corresponding papers and publicly available implementations, including optimization settings, training schedules, and data augmentation strategies. For SACFusion, we follow the standard DINO augmentation pipeline, including multi scale resizing and the random crop branch. All methods are evaluated on the same dataset splits using the same COCO style evaluation protocol. Model selection is based on validation performance, and no test time augmentation is used unless required by the original method.

For DSEC-Det, all methods use the original input resolution of 640×480. For PEOD, whose native resolution is 1280×720, all inputs are uniformly resized to 640×360, since several compared methods are difficult to train reliably at the original resolution under the same hardware setting. This unified resolution ensures consistent computational conditions across methods.

B Additional Quantitative Results

B.1 Moderate Illumination Conditions

Qualitative results under moderate illumination are shown in Fig. 7. In these scenes, RGB appearance remains relatively reliable, providing a strong basis for detection. Nevertheless, small, distant, partially occluded, or weakly contrasted objects can still be missed by YOLOX-M, especially in crowded or structurally complex regions. MvHeat-DET provides additional motion and contour cues from Event observations, but sparse responses may lead to incomplete object structures or unstable localization. SPFD benefits from multimodal fusion, yet several difficult targets remain missed or imprecisely localized. In comparison, SACFusion produces predictions that are more consistent with the ground truth across the seven examples. These results indicate that the proposed framework can exploit complementary Event evidence even when RGB observations are already informative, while avoiding excessive interference with stable RGB representations.



Figure 7: Qualitative comparison under moderate illumination conditions on DSEC-Det and PEOD. From left to right, results from YOLOX-M, MvHeat-DET, SPFD, SACFusion, and ground truth (GT) are shown. The examples cover diverse urban scenes in which RGB appearance remains relatively informative.

B.2 Challenging Illumination Conditions

Under challenging illumination, RGB appearance becomes severely degraded, particularly for distant and low contrast targets, as illustrated in Fig. 8. In the upper three examples, YOLOX-M can still detect several visually salient vehicles but tends to miss small or weakly visible targets in darker regions. MvHeat-DET benefits from the clearer motion and contour structures preserved by Event observations, although its predictions are less stable and occasionally incomplete. SPFD recovers part of the missing information through multimodal fusion, but several distant targets remain undetected. In comparison, SACFusion produces predictions that are more consistent with the ground truth, especially for multiple distant vehicles and other weakly visible traffic participants. These examples show that the proposed semantic compatibility learning and asymmetric Event retrieval can make better use of complementary Event evidence when RGB appearance becomes unreliable under severe illumination degradation.

B.3 High Speed Motion Conditions

High speed motion introduces strong blur and rapidly changing object structures, making appearance based detection more difficult, as shown in Fig. 8. YOLOX-M often misses fast moving or small objects whose RGB appearance becomes ambiguous, while MvHeat-DET captures motion related responses but may produce unstable localization due to sparse observations. SPFD improves

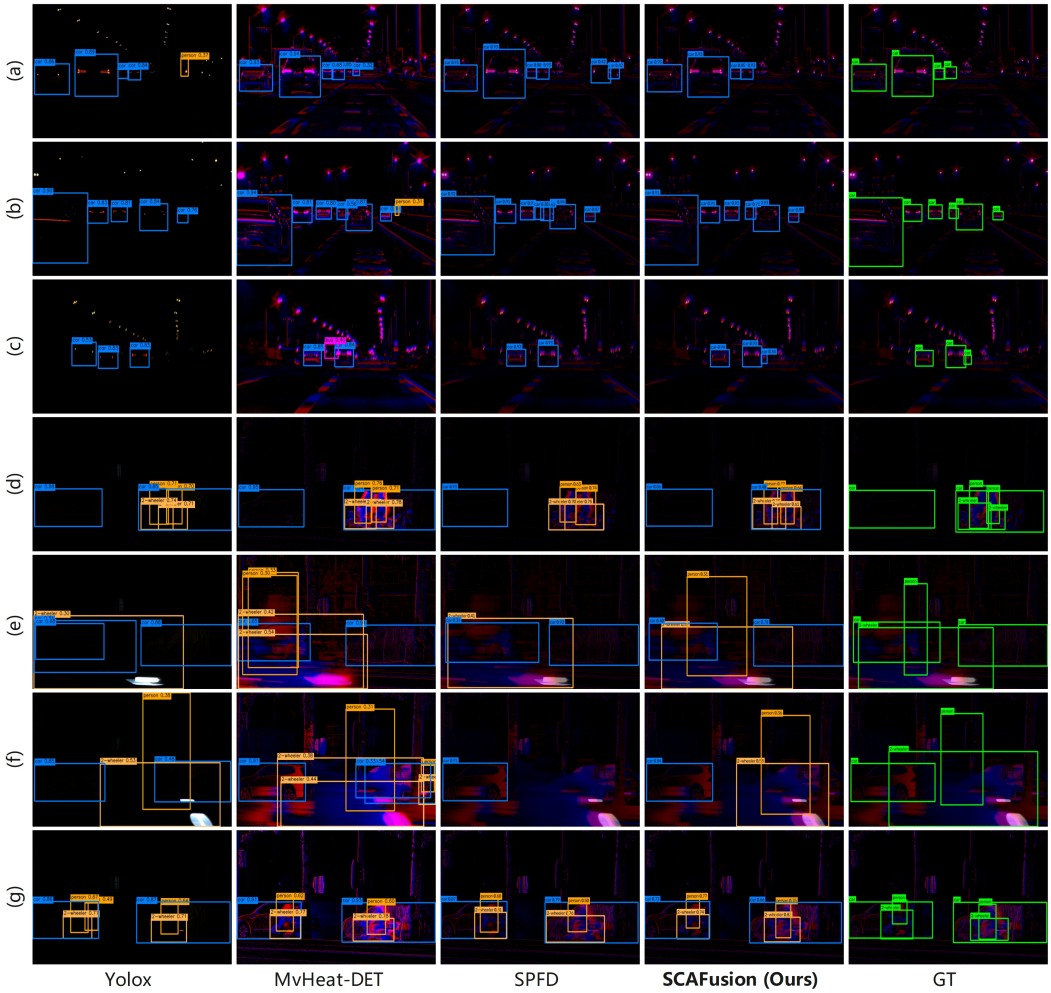

Figure 8: Qualitative comparison under challenging illumination and high speed motion conditions on DSEC-Det and PEOD. Rows (a)-(c) show scenes with challenging illumination, while rows (d)-(g) show scenes with high speed motion. From left to right, results from YOLOX-M, MvHeat-DET, SPFD, SACFusion, and ground truth (GT) are shown.

multimodal detection, but some moving targets remain difficult to localize accurately. SACFusion achieves more reliable predictions by selectively retrieving complementary Event evidence from informative regions while retaining stable RGB semantics. The results suggest that Event information provides particularly useful dynamic cues under rapid motion, and that the proposed asymmetric retrieval mechanism can exploit these cues more effectively for robust detection.

C Supplementary Feature Response Analysis

Additional qualitative evidence of semantic compatibility is provided through feature response visualization across different backbone levels, as shown in Fig. 9. In the first example, both RGB and Event inputs contain clear target information. As the representations progress from res3 to res5, the high response regions of the two modalities become increasingly consistent around semantically relevant objects, illustrating effective cross modal alignment when both observations are informative.

The remaining examples further demonstrate that semantic compatibility does not require equally informative inputs from both modalities. In the second example, RGB contains substantially richer target cues than Event, yet the deeper RGB and Event features still attend to similar object regions at res5. In the third example, Event observations provide more informative evidence for the vehicle on the right, and this target remains clearly emphasized in the deeper Event representation. These observations suggest that the proposed alignment encourages semantic consistency without forcing

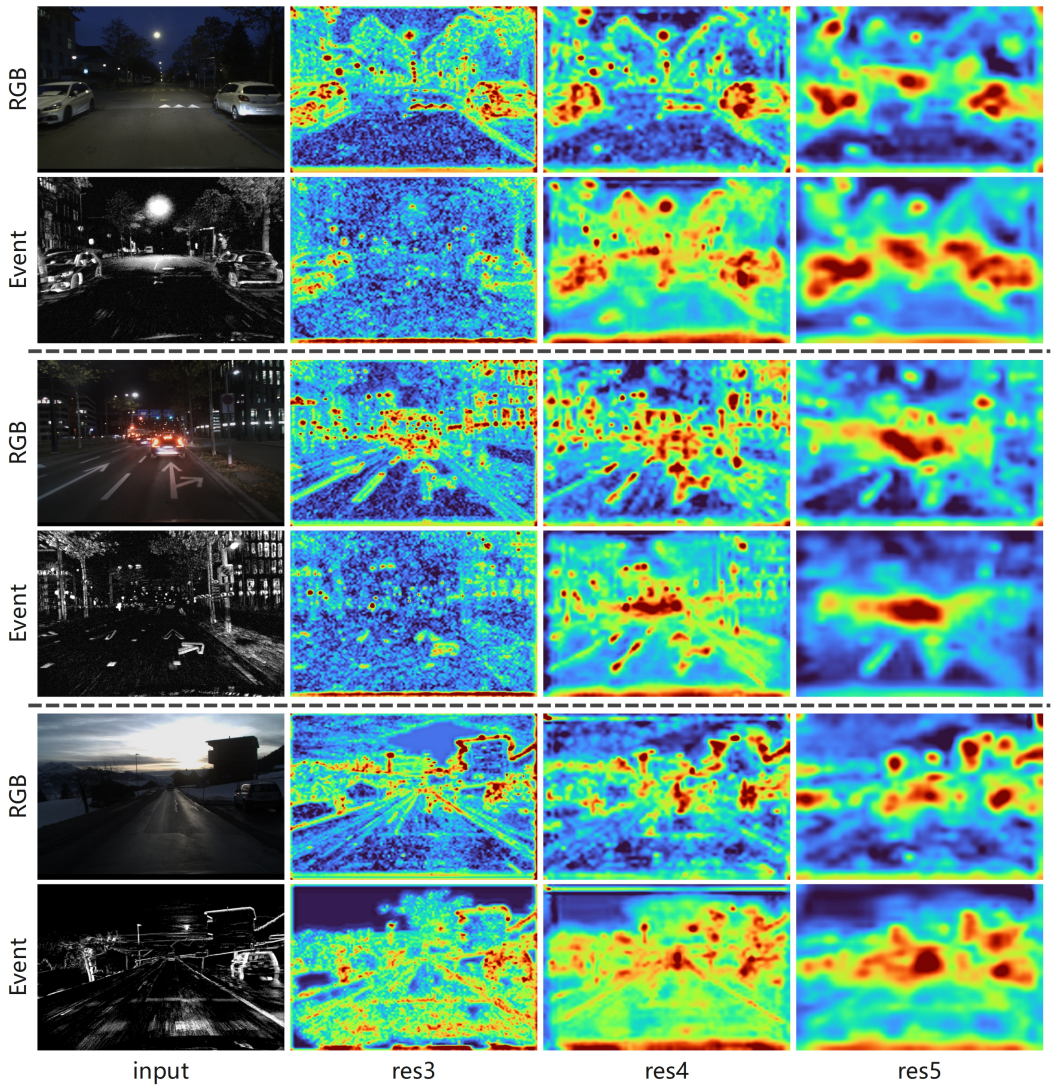

Figure 9: Supplementary visualization of RGB and Event feature responses across backbone levels. The three examples illustrate different cross modal conditions. The first pair contains informative observations in both modalities, the second pair has richer RGB target cues but relatively weak Event observations, and the third pair contains more informative Event evidence for the vehicle on the right. Feature responses at res3, res4, and res5 are visualized for comparison.

the two modalities to become identical, thereby preserving modality specific evidence for subsequent complementary fusion.

D Discussion

Although SACFusion improves RGB–Event detection under a wide range of challenging conditions, several difficult cases remain, as illustrated in Fig. 10. Small and distant objects are still prone to missed detections when their appearance and Event responses are both weak. Partial occlusion, weak object boundaries, and limited visible structure can also lead to inaccurate localization or classification. These cases suggest that multimodal fusion cannot fully compensate for situations in which neither modality provides sufficiently discriminative evidence.

More broadly, the effectiveness of SACFusion still depends on the quality and complementarity of the available RGB and Event observations. The proposed semantic compatibility learning and asymmetric retrieval reduce modality bias, but they do not eliminate ambiguity caused by intrinsically

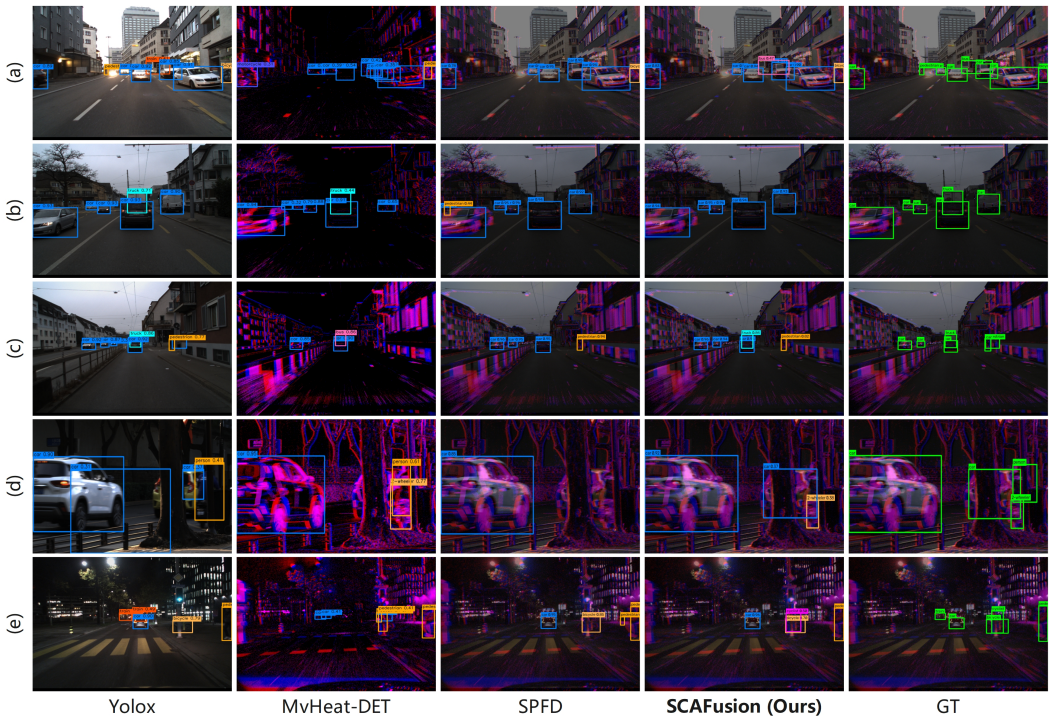

Figure 10: Representative challenging cases on DSEC-Det and PEOD. From left to right, results from YOLOX-M, MvHeat-DET, SPFD, SACFusion, and ground truth (GT) are shown.

weak observations. Future work could therefore investigate more explicit reliability or uncertainty modeling to better estimate when complementary evidence is trustworthy, as well as stronger temporal reasoning to exploit information beyond a single synchronized observation. Such extensions may further improve robustness in difficult cases where spatial evidence from both modalities is limited.