RETDET: DEGRADATION-AWARE TEXT AS A DETACHABLE PRIOR FOR RGB-EVENT OBJECT DETECTION

Anonymous authors

Paper under double-blind review

ABSTRACT

By fusing RGB images and event data, multimodal object detection improves accuracy and robustness in complex scenes, yet attention-dominated fusion modules incur high computational cost, and under degraded conditions the representation capability of both modalities declines. High-level semantic priors describing the degradation itself could compensate for this loss, but querying a multimodal large language model (MLLM) at every inference step is prohibitively expensive. We propose RETDet, an RGB-Event-Text detection framework in which degradation-aware text acts as a detachable prior. During training, MLLM-generated degradation descriptions are mapped into layer-wise modulation parameters and randomly replaced by a learned null embedding, so that the detector internalizes degradation priors into its visual pathway while remaining able to consume explicit descriptions. The trained model supports two inference modes: a prior mode taking only RGB and event inputs, invoking neither an MLLM nor a text encoder at deployment, and a conditioned mode accepting scene descriptions for higher accuracy. Without any textual input at inference, prior mode already outperforms SOTA RGB-Event methods on DSEC, PKU-DDD17-Car, TUMTraf, and EventKITTI, and conditioned mode yields further gains concentrated on the most degraded subsets, decoupling the semantic benefit of MLLM supervision from its inference cost.

1 INTRODUCTION

Object detection is a fundamental task in computer vision (Chen et al., 2021; Oksuz et al., 2020). In autonomous driving perception systems, accurate recognition of pedestrians, vehicles, traffic signs, and obstacles is crucial for safe and reliable operation (Ma et al., 2020). In recent years, deep-learning-based object detection has achieved remarkable progress (Liu et al., 2020; Redmon et al., 2016; Lin et al., 2017b; Ren et al., 2015; Cai & Vasconcelos, 2018; Zhu et al., 2021); however, in autonomous driving scenarios, RGB images are easily affected by degradation factors such as motion blur, overexposure, and low illumination, leading to limited detection performance. To address this issue, bio-inspired event cameras have been introduced as a complementary visual sensing modality (Brandli et al., 2014; Moeys et al., 2017). By asynchronously recording pixel-wise brightness changes (Chen et al., 2020; Gallego et al., 2020), event cameras provide high temporal resolution, high dynamic range, and low latency, thereby effectively compensating for the limitations of RGB images in dynamic information capture and adaptation to extreme lighting conditions. Therefore, recent studies have increasingly explored RGB-Event fusion methods and achieved significant progress.

Representative fusion modules such as FlexEvent (Lu et al., 2025) and CAFR (Cao et al., 2024) typically rely on convolutional layers or attention mechanisms to enhance cross-modal interaction. Although such designs improve feature representation capability, they also bring higher model complexity and computational cost; the cross-attention operation in CAFR, for example, grows quadratically with feature-map resolution, i.e., $O(N^2)$. In addition, current methods mostly focus on feature-level interaction between visual modalities, but lack explicit modeling of scene degradation conditions. Since both RGB and event data may be affected by various degradation factors, the effectiveness of information from each modality dynamically fluctuates with environmental changes. Fusion mechanisms that rely only on visual data are therefore difficult to regulate adaptively and cannot effectively compensate for the information loss caused by degradation.

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.

A natural way to supply the missing environmental context is to describe it in language. Structured textual descriptions can jointly express the type and severity of multiple degradation factors in an explicit form, and such descriptions can be produced by an MLLM. This, however, creates a practical tension. Obtaining a scene-specific description at every inference step usually requires an online MLLM query, which in our measurements adds more than one second of latency per frame. The central question of this work is therefore not only whether textual degradation priors help, but whether their benefit can be retained without paying their inference cost.

Our answer is to treat degradation-aware text as a detachable prior rather than as a mandatory input. During training, MLLM-generated descriptions are encoded and mapped into layer-wise modulation parameters that reshape the visual features, and are randomly replaced by a learned null embedding. Textual supervision thus shapes the visual pathway throughout training, while the network is simultaneously trained to operate without any description. At inference, a single model exposes two modes that differ only in where the text embedding comes from. In prior mode, the optimized null embedding is used, so neither an MLLM nor a text encoder is executed, and the inference cost is identical to a purely visual RGB-Event detector. In conditioned mode, an explicit description is supplied, generated online or retrieved from cache.

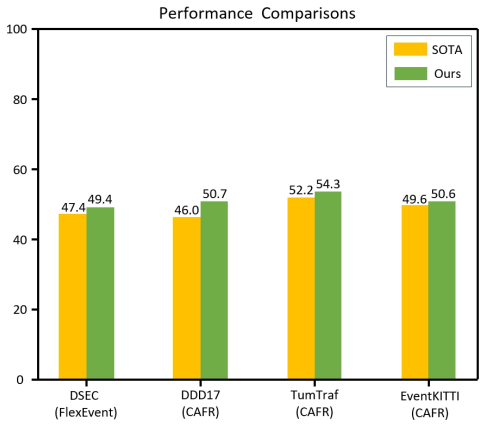


Figure 1: Performance comparison on four public datasets. For each dataset we report the previous SOTA method and our prior mode performance, which requires no textual input at inference. Prior mode already surpasses the previous SOTA on all four datasets.

Building on this formulation, we propose RETDet. Compared with existing high-complexity fusion designs, we introduce a Parameter-Free Enhanced Module (PFEM) as the core of the fusion network, which enables cross-modal feature interaction and alignment without any additional learnable parameters. To incorporate the textual modality, we employ Qwen-VL-Max (Bai et al., 2023) to generate structured descriptions of scene degradation, which a Textual Modulation Module (TMM) maps into per-layer modulation parameters. The Text-Guided Bayesian Refinement Module (TBRM) applies these parameters along a pyramid coupled to each FPN (Lin et al., 2017a) level, so that degradation priors reach features at every scale, and then treats highly overlapping proposals as repeated noisy observations of one latent box, recovering it by confidence-weighted maximum a posteriori estimation rather than discarding them through NMS (Hosang et al., 2017). As shown in Fig. 1, prior mode already outperforms existing SOTA approaches on DSEC (Gehrig et al., 2021b), PKU-DDD17-Car (Li et al., 2019), TUMTraf (Creß et al., 2024), and EventKITTI (Liang et al., 2022) without consuming any text at inference.

The main contributions of this work can be summarized as follows:

- We formulate degradation-aware text as a detachable prior for RGB-Event detection. Textual supervision shapes the visual pathway during training and becomes optional at inference, which decouples the semantic benefit of MLLM guidance from its inference cost.
- We instantiate this formulation as a single model with two inference modes. Prior mode requires no MLLM and no text encoder and matches the deployment cost of a purely visual detector, while conditioned mode consumes explicit scene descriptions for higher accuracy.
- We propose an RGB-Event Parameter-Free Fusion Network centered on PFEM, which performs cross-modal enhancement and statistical alignment without additional learnable parameters, and a Text-Guided Proposal Refinement Network centered on TBRM, which couples textual modulation with multi-scale visual features and aggregates overlapping proposals as repeated noisy observations of a latent object.
- Without any textual input at inference, our method consistently outperforms SOTA RGB-Event approaches on DSEC, PKU-DDD17-Car, TUMTraf, and EventKITTI, while also running faster than all compared fusion methods.

2 Related Work

Event-Based Object Detection. Event-based object detection methods fall into two categories. The first converts event streams into dense representations, including event frames or voxel grids, enabling conventional deep learning architectures. Representative methods include RED (Perot et al., 2020), which uses ConvLSTM (Hochreiter & Schmidhuber, 1997) for spatiotemporal aggregation, ASTMNet (Li et al., 2022a), which introduces spatiotemporal attention and a lightweight memory module, and RVT (Gehrig & Scaramuzza, 2023), which employs a recurrent vision transformer to capture long-range temporal dependencies. The second exploits the sparsity and asynchrony of events, including GNN-based methods (Schaefer et al., 2022), SNN-based methods (Cordone et al., 2022), and hybrid ANN-SNN architectures (Wang et al., 2023). Although these event-native methods better preserve event-stream properties, their detection accuracy and hardware support remain limited. This work therefore follows the dense-representation paradigm.

RGB-Event Fusion Module. Multimodal event-frame fusion has been widely explored in vision tasks. In image deblurring, EFNet (Sun et al., 2022) introduces an Event-Image Cross-modal Attention module to combine motion cues with spatial details from RGB frames. In semantic segmentation, CMX (Zhang et al., 2023) improves feature consistency across heterogeneous modalities via cross-modal rectification and fusion. In object detection, CAFR (Cao et al., 2024) employs bidirectional cross-modality interaction and adaptive feature refinement to align RGB and event representations in challenging scenes, while FlexEvent (Lu et al., 2025) introduces adaptive event-frame fusion for arbitrary-frequency object detection. Despite these advances, fusion methods face two limitations: high computational complexity in attention-based modules, and limited explicit modeling of environmental conditions. Under scene degradation, models cannot reason about the sensing environment and instead rely on data-driven adjustment of modality contributions. We introduce textual information as auxiliary guidance to characterize degradation conditions.

Text-Guided Visual Perception. With the development of vision-language pre-training, textual information has been widely introduced into vision tasks to enhance scene understanding. CLIP (Radford et al., 2021) establishes general vision-language alignment through large-scale image-text contrastive learning, while BLIP (Li et al., 2022b) further improves cross-modal representation learning. In downstream tasks, text has been applied to object detection, semantic segmentation, and open-vocabulary perception. For example, GLIP (Li et al., 2022c) introduces language supervision into object detection and learns language-aware object-level visual representations. In text-guided segmentation, Yi et al. (2023) and Wu et al. (2024) demonstrate the effectiveness of text supervision and image-text co-decomposition, respectively. Two properties distinguish our use of language from this line of work. First, text describes the sensing conditions rather than object categories or regions, which is what allows a small, structured vocabulary of degradation types to cover a scene. Second, because the description characterizes conditions rather than content, it is meaningful to withhold it. A category prompt is indispensable to an open-vocabulary detector, whereas a degradation description can be replaced by a learned null embedding once its effect has been absorbed into the visual pathway during training.

3 Methodology

In this section, we present RETDet in detail. Sec. 3.1 introduces the RGB-Event Parameter-Free Fusion Network, which enables efficient feature fusion between the RGB and event visual modalities. Sec. 3.2 describes the Text-Guided Proposal Refinement Network, which uses the textual modality to modulate visual features and refine fused region features. Sec. 3.3 then formalizes the two inference modes and the training procedure that makes both modes available from a single model. The overall framework is illustrated in Fig. 2.

3.1 RGB-Event Parameter-Free Fusion Network

RGB-Event Feature Extraction. We extract multi-scale RGB features with ResNet-50 (He et al., 2016) and event features with the Recurrent Vision Transformer (RVT) (Gehrig & Scaramuzza, 2023), which combines spatial attention with recurrent temporal processing. The two feature hierarchies are then fused by PFEM.

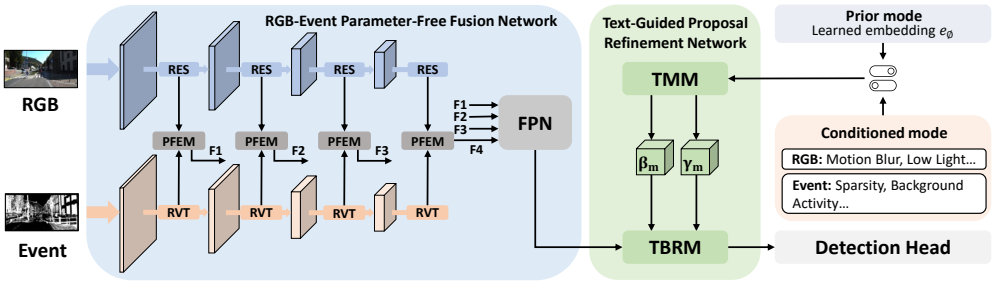


Figure 2: RETDet architecture. PFEM fuses RGB-event features, and TMM maps either a description embedding or the learned null embedding $e_\varnothing$ to layer-wise modulation parameters. TBRM refines multi-scale features and proposals for detection. Prior and conditioned modes share all weights and differ only in the active embedding.

Parameter-Free Enhanced Module. The fusion module is essential for exploiting complementary information across multiple modalities, yet existing methods are often architecturally complex and computationally expensive, which limits real-time inference. To address this issue, we design a Parameter-Free Enhanced Module (PFEM), as illustrated in Fig. 3. In the initial stage, PFEM enhances the input features through a parameter-free attention mechanism based on neuron energy modeling. This mechanism is inspired by SimAM (Yang et al., 2021) and characterizes neuron importance by evaluating redundancy in the feature representation. Specifically, the importance of each neuron is determined by its corresponding energy function, where a lower energy value indicates higher importance. Thus, the importance can be expressed as $1/e_n^*$.

$$e_n^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(n - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{1}$$

where $\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$, $\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{\mu})^2$, $M = H \times W$ denotes the number of neurons in that channel, the target neuron n lies in a single channel of the input feature $X \in \mathbb{R}^{C\times H\times W}$, and λ is a regularization term used to stabilize the energy estimation. The feature is then adaptively enhanced through the following scaling operation:

$$\tilde{X} = \text{sigmoid}\left(\frac{1}{E}\right) \odot X \tag{2}$$

where E aggregates all e_n^* across channel and spatial dimensions. Furthermore, considering the large representational gap between event data and RGB images, PFEM introduces a parameter-free adaptation mechanism based on statistical alignment during cross-modal fusion. Inspired by AdaIN (Huang & Belongie, 2017), we introduce a parameter-free statistical alignment operation. RGB and event features serve as content and style representations. By aligning the channel-wise mean and variance of the two modalities, the RGB features are shifted toward the event feature distribution, thereby improving cross-modal consistency and fusion capability. This process is formulated as:

$$\text{StatAlign}(c, s) = \sigma(s)\left(\frac{c - \mu(c)}{\sigma(c)}\right) + \mu(s) \tag{3}$$

where c and s denote the content input and style input, respectively. Through this design, PFEM enhances and aligns multimodal features without introducing any additional learnable parameters. A component-wise analysis of PFEM is provided in Sec. 4.4. A Feature Pyramid Network (FPN) (Lin et al., 2017a) combines the PFEM outputs through top-down and lateral connections to produce multi-scale RGB-event features for subsequent refinement.

3.2 Text-Guided Proposal Refinement Network

To introduce degradation-aware textual priors, we employ the vision-language model Qwen-VL-Max (Bai et al., 2023) to jointly analyze the quality of RGB and event images and generate structured semantic descriptions of degradation conditions. Descriptions are generated offline for the training set and stored alongside the corresponding RGB-event sample, so the MLLM is not queried repeatedly during training. Details of the generation process and the prompt design are provided in **Appendix A**.

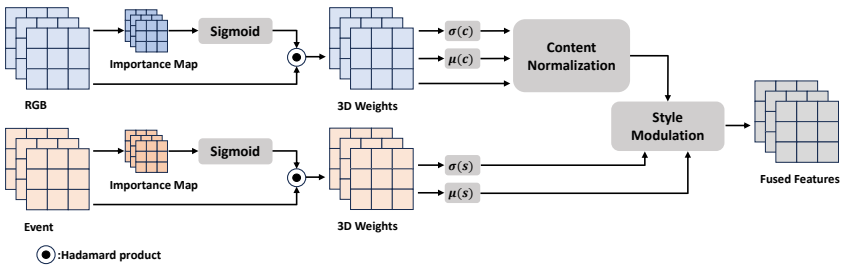


Figure 3: Structure of PFEM. RGB and Event features are enhanced via importance-aware weighting and aligned through content normalization and style modulation to produce fused features.

Textual Modulation Module. A CLIP-based (Radford et al., 2021) Textual Modulation Module (TMM) encodes structured textual descriptions into semantic representations that can interact with visual features, as illustrated in Fig. 4. TMM first adopts a BPE tokenizer (Sennrich et al., 2016) to convert a description into a token sequence acceptable to CLIP; the sequence is then fed into the CLIP text encoder, where tokens are mapped into the embedding space, positional encodings are added, and multiple Transformer layers model contextual dependencies, producing a fixed-dimensional global textual feature vector $e_{\text{txt}} \in \mathbb{R}^d$.

In addition to e_{txt}, TMM maintains a learned null embedding $e_{\varnothing} \in \mathbb{R}^d$, a free parameter of the same dimensionality that is optimized jointly with the rest of the network. It represents the absence of an explicit description and, as described in Sec. 3.3, is exercised throughout training rather than introduced only at deployment. We write the active embedding as

$$e = \begin{cases} e_{\text{txt}}, & \text{conditioned mode,} \\ e_{\varnothing}, & \text{prior mode.} \end{cases} \tag{4}$$

The active embedding is fed into MLP_m corresponding to each FPN layer, where $m \in \{2, 3, 4, 5\}$. For the m-th layer, MLP_m outputs two modulation parameters, γ_m and β_m, corresponding to the scaling and shifting terms used to modulate the visual features at that layer. Benefiting from the cross-modal alignment acquired by CLIP through image-text joint training, a semantic correspondence can be established between textual and visual features in a progressive manner, improving the effectiveness of textual priors in modulating visual representations.

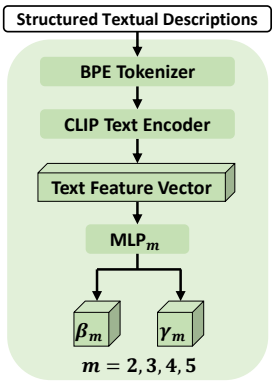


Figure 4: Structure of TMM. A structured description is processed by the BPE tokenizer and CLIP text encoder to obtain a textual feature vector, which is fed into layer-wise MLPs to generate modulation parameters. In prior mode this vector is replaced by the learned null embedding $e_{\varnothing}$.

Text-Guided Bayesian Refinement Module. To achieve hierarchical semantic fusion between textual information and multi-scale visual features while improving region representations, we propose the Text-Guided Bayesian Refinement Module (TBRM). As illustrated in Fig. 5, TBRM employs the Text-Guided Pyramid Network (TPN) to progressively enhance the multi-scale features from the FPN. Starting from the top-level feature map P_5, a spatial attention module first highlights informative visual regions, such as object boundaries and texture-rich areas. The resulting feature is then modulated by the corresponding parameters γ_m and β_m, and further processed by a self-attention module to obtain the updated representation $T\text{-}P_5$. This feature is upsampled to the same resolution as P_4, modulated by the corresponding parameters, concatenated with P_4, and passed through a 1×1 convolution for channel alignment. Another self-attention module is then applied to generate $T\text{-}P_4$. This top-down process is repeated along the pyramid, producing a complete multi-scale feature representation. Visualized feature maps are provided in Sec. 4.1.

The Region Proposal Network (RPN) (Ren et al., 2015) produces proposals $\{b_i\}_{i=1}^N$ with objectness scores $\{s_i\}_{i=1}^N$. We connect proposals whose pairwise IoU exceeds τ and form clusters $\{C_k\}$ by connected component analysis. Instead of suppressing overlapping proposals through NMS (Hosang et al., 2017), we treat each proposal in a cluster as a noisy observation of a latent box $b^{(k)}$:

$$b_i = b^{(k)} + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2 \mathbf{I}), \qquad \sigma_i^2 \propto 1/s_i. \tag{5}$$

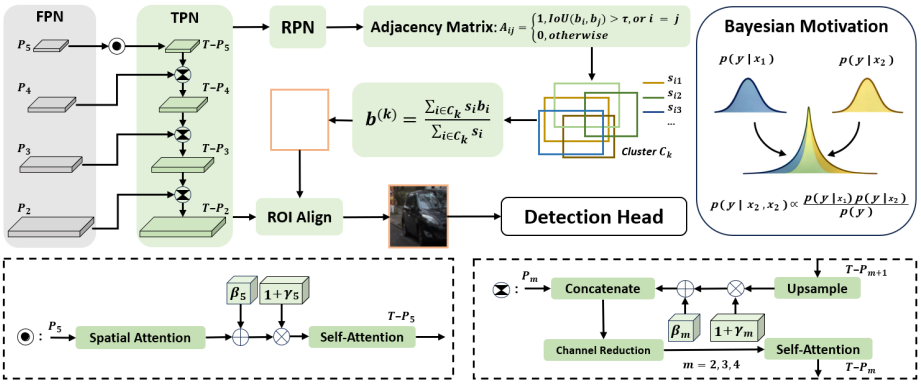


Figure 5: Architecture of the proposed Text-Guided Bayesian Refinement Module (TBRM). TBRM fuses the textual semantics with multi-scale FPN features via TPN, refines RPN proposals through confidence-weighted aggregation, and extracts region features using RoI Align for final detection.

where the objectness score s_i is taken as a proxy for observation reliability. Under a flat prior and conditionally independent observations, the posterior is Gaussian, and its mean is the precision-weighted average of the observations. Inspired by Weighted Boxes Fusion (WBF) (Solovyev et al., 2021), it reduces exactly to:

$$b^{(k)} = \frac{\sum_{i \in C_k} \sigma_i^{-2} b_i}{\sum_{i \in C_k} \sigma_i^{-2}} = \frac{\sum_{i \in C_k} s_i b_i}{\sum_{i \in C_k} s_i}. \tag{6}$$

Without additional learnable parameters, this strategy integrates multiple spatial observations of the same object, reduces redundant detections, and preserves localization information, yielding more precise proposals. After refinement, proposals are normalized to $[0, 1]$, and the top-M proposals with the highest scores are selected and fed into RoI Align (He et al., 2017) for region feature extraction. RoI Align assigns each proposal to the appropriate TPN-generated feature layer according to its scale:

$$x = x_0 + \log_2 \left(\frac{\sqrt{wh}}{s_{\text{ref}}} \right) \tag{7}$$

where w and h denote the width and height of the proposal, s_{ref} denotes the reference scale determined by the typical object size in the dataset, and x_0 is the base pyramid layer. Proposals of different scales are thus converted into fixed-size feature representations and fed into the detection head for classification and bounding box regression.

3.3 Training with Text Dropout and Two Inference Modes

Both operating regimes are learned in one training run through text dropout. The null embedding is trained jointly rather than introduced only at deployment. At each training step, the active embedding in Eq. 4 is set to $e_\varnothing$ with probability p and to e_{txt} otherwise, sampled independently per sample, with $e_\varnothing$ optimized by the detection loss alongside all other parameters. This is the conditioning-dropout strategy of classifier-free guidance (Ho & Salimans, 2022), applied to a degradation prior rather than a generative condition. Its effect is twofold: the network cannot place the entire degradation-handling burden on the text branch, since that branch carries no scene information on a fraction p of samples; and prior mode is explicitly optimized during training rather than introduced as an untrained fallback.

At inference, the two modes share all weights and modules and differ only in the active embedding. Prior mode uses $e_\varnothing$, which is constant after training. Its modulation parameters $\{\gamma_m^\varnothing, \beta_m^\varnothing\}$ are precomputed once and stored as buffers, so neither the MLLM nor the CLIP text encoder is instantiated at deployment and the text branch contributes zero runtime. The inference cost is that of a purely visual RGB-Event detector. Conditioned mode uses e_{txt}, obtained by encoding a description generated online, cached offline, retrieved from a small bank of preconfigured descriptions covering expected deployment conditions, or specified manually when conditions are known. Only online generation requires an MLLM at inference. The textual prior is therefore detachable: MLLM guidance is acquired during training, while its inference-time use remains optional.

Table 1: Comparison with RGB-Event fusion methods (mAP@0.50:0.95). All methods use RGB-Event inputs at inference; ours runs in prior mode without an MLLM or text encoder. FlexEvent is only evaluated on DSEC. Best in **bold**, second best underlined.

Method	DSEC				PKU-DDD17-Car		
	Car	Pedestrian	Large veh.	Average	Day	Night	All
CMX T-ITS'23	0.416	0.164	0.294	0.291	0.402	0.354	0.390
REFusion ICRA'23	0.405	0.172	0.306	0.294	0.454	0.391	0.439
FPNFusion ICRA'22	0.375	0.109	0.249	0.244	0.432	0.357	0.416
FAGC IEEE Sens. J.'21	0.398	0.144	0.336	0.293	0.448	0.395	0.436
EFNet ECCV'22	0.411	0.158	0.326	0.298	0.434	0.351	0.416
RAMNet RA-L'21	0.244	0.108	0.176	0.176	0.392	0.369	0.388
CAFR ECCV'24	0.499	0.258	0.382	0.380	0.469	0.421	0.460
FlexEvent NeurIPS'25	0.593	0.374	**0.455**	0.474	—	—	—
Ours	**0.632**	**0.417**	0.432	**0.494**	**0.517**	**0.463**	**0.507**

Method	TUMTraf						EventKITTI			
	Ped.	Car	Bus	Truck	Trailer	Avg.	Car	Ped.	Cyclist	Avg.
CMX T-ITS'23	0.323	0.762	0.550	0.218	0.342	0.439	0.681	0.306	0.435	0.474
REFusion ICRA'23	0.392	0.775	0.603	0.244	0.476	0.498	0.694	0.328	0.468	0.495
FPNFusion ICRA'22	0.248	0.767	0.544	0.297	0.454	0.462	0.680	0.330	0.468	0.493
FAGC IEEE Sens. J.'21	0.418	0.773	0.710	0.198	0.443	0.508	**0.705**	0.337	0.449	0.496
EFNet ECCV'22	0.285	**0.776**	0.664	0.291	**0.500**	0.503	0.689	0.311	0.440	0.480
RAMNet RA-L'21	0.345	0.695	0.483	0.225	0.238	0.397	0.472	0.179	0.353	0.335
CAFR ECCV'24	0.335	0.772	**0.764**	0.255	0.484	0.522	0.697	0.316	0.475	0.496
Ours	**0.423**	0.754	0.713	**0.339**	0.487	**0.543**	0.667	**0.360**	**0.492**	**0.506**

4 EXPERIMENTS

The central claim of RETDet is that degradation-aware text can act as a detachable prior. We evaluate whether text-derived benefits persist without descriptions at inference (Sec. 4.1), quantify the cost of each inference mode (Sec. 4.2), and assess the additional benefit of explicit descriptions (Sec. 4.3). Additional ablation studies are introduced in Sec. 4.4, and implementation details are given in **Appendix B.4**.

4.1 PRIOR MODE: COMPARISON WITH STATE-OF-THE-ART WITHOUT TEXTUAL INPUT

All results in this subsection are obtained in prior mode, where the text branch consumes the learned null embedding, and neither an MLLM nor a text encoder is invoked at inference. Our model therefore takes exactly the same inputs as the RGB-Event baselines, which consume no textual signal. Conditioned-mode results, where an explicit scene description is supplied, are reported separately in Sec. 4.3.

Since existing studies lack investigations on introducing the textual modality into RGB-Event fusion-based detection, we compare with representative RGB-Event fusion approaches, including CMX (Zhang et al., 2023), REFusion (Zhou et al., 2023), FPNFusion (Tomy et al., 2022), FAGC (Cao et al., 2021), EFNet (Sun et al., 2022), RAMNet (Gehrig et al., 2021a), and CAFR (Cao et al., 2024). On DSEC we additionally include FlexEvent (Lu et al., 2025). Baseline results follow the protocol of Cao et al. (2024), in which each fusion module is embedded in RetinaNet; FlexEvent results are taken from Lu et al. (2025). Detailed dataset descriptions are in **Appendix B.1**, and qualitative comparisons in **Appendix C.1**.

Across all four benchmarks, prior mode achieves the best overall mAP@0.50:0.95 without consuming any textual input at inference: 0.494 on DSEC against 0.474 for FlexEvent (Lu et al., 2025), and 0.507 on PKU-DDD17-Car against 0.460 for CAFR (Cao et al., 2024), leading in both the day and night splits; 0.543 on TUMTraf against 0.522, and 0.506 against 0.496 on EventKITTI (Table 1). Category-wise results also highlight strong performance on smaller road users, particularly pedestrians on DSEC and EventKITTI and cyclists on EventKITTI.

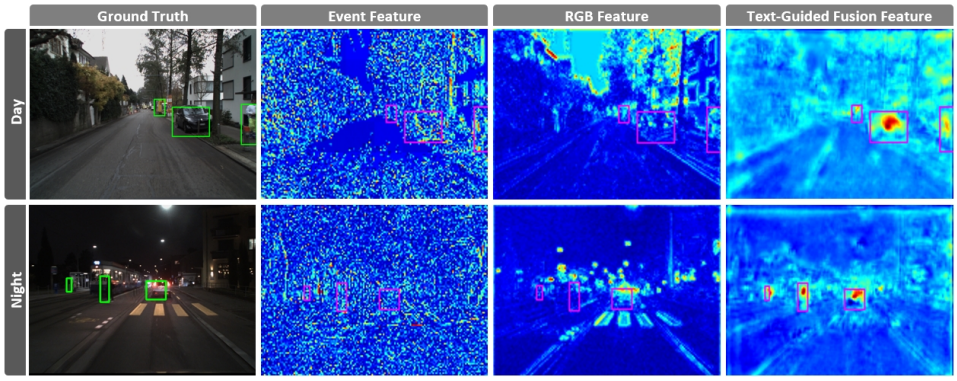


Figure 6: Prior-mode feature maps in daytime and nighttime scenes. From left to right: ground truth, event features, RGB features, and fused features after TBRM with the learned null embedding.

Because prior mode feeds a constant embedding, these margins must originate in training rather than in test-time guidance, and Table 2 probes where. Removing the text modulation branch (TPN) and retraining lowers TUMTraf mAP from 0.543 to 0.523, while replacing the learned null embedding with a random one of the same dimensionality gives 0.527. The benefit therefore depends on the embedding the network was optimized around, not merely on the presence of a conditioning pathway or additional parameters. Fig. 6 visualizes this effect. Although no description is supplied, the fused prior-mode features respond more strongly on targets and suppress background clutter compared with either single-modality feature, including the noisy event responses and the dark RGB regions at night.

Table 2: Prior-mode ablation on TUMTraf (mAP@0.50:0.95).

Configuration	Text supervision	mAP
No TPN (retrained)	✗	0.523
Random embedding	✓	0.527
Prior mode embedding	✓	**0.543**

Table 3: Inference cost. (a) Per-image cost of the two modes on TUMTraf (640×480). Detector time includes proposal IoU computation, clustering, and fusion. (b) Speed and accuracy against the RGB-Event fusion baselines on PKU-DDD17-Car (346×260). mAP denotes mAP@0.50:0.95.

(a) Inference modes on TUMTraf

Mode	Text source	MLLM (ms)	CLIP (ms)	Detector (ms)	Total (ms)	FPS
Prior	precomputed $e_\varnothing$	—	—	306.6	**306.6**	**3.26**
Conditioned	cache / preset bank	—	2.6	306.6	309.2	3.23
Conditioned	Qwen-VL-Max, online	1023.0	2.6	306.6	1332.2	0.75

(b) Efficiency comparison with fusion baselines on PKU-DDD17-Car

	CMX	REFusion	FAGC	CAFR	EFNet	RAMNet	FPNFusion	Ours (prior)
FPS	2.80	5.01	5.27	6.39	9.71	11.45	12.10	**13.0**
mAP	0.390	0.439	0.436	0.460	0.416	0.388	0.416	**0.507**

4.2 Efficiency and Deployment Cost

Table 3(a) reports the per-image cost of both modes. Because the modulation parameters derived from $e_\varnothing$ are precomputed, the text branch adds no runtime in prior mode. Cached or preset descriptions recover the full conditioned accuracy for 2.6 ms, which suits deployments whose conditions are known in advance or fall within a small enumerable set. Genuinely unfamiliar scenes still require online generation at over 1 s per frame. The model contains 98.147M parameters and requires 425.142G FLOPs in prior mode on TUMTraf. Against the compared fusion methods on PKU-DDD17-Car at 346×260 (Table 3(b)), prior mode is both the fastest and the most accurate: it runs at 13.0 FPS, slightly ahead of the lightest baseline FPNFusion (12.10 FPS), and about twice as fast as the strongest baseline CAFR (6.39 FPS) while raising mAP from 0.460 to 0.507.

4.3 Conditioned Mode

Table 4 reports both modes on the three datasets for which we generated scene-specific descriptions during inference. Supplying an explicit description improves accuracy on all three. Read together with Table 2, where a random embedding of the same dimensionality falls below the null embedding, this establishes that accuracy tracks the semantic content of the conditioning signal rather than its presence. Although conditioned mode provides additional accuracy gains, we recommend prior mode as the default for deployment because it eliminates inference-time description generation and text encoding while retaining strong detection performance. We also ablate the conditioning source by replacing the text embedding with a CLIP image embedding in **Appendix C.2**.

Table 4: Conditioned (Cond.) vs. prior mode, mAP@0.50:0.95 on full test sets. The last row keeps the 314 TUMTraf frames whose description reports moderate or higher degradation.

Evaluation set	Prior	Cond.	Δ
PKU-DDD17-Car	0.507	0.509	+0.002
EventKITTI	0.506	0.509	+0.003
TUMTraf	0.543	0.552	+0.009
TUMTraf (degraded)	0.539	0.551	+0.012

The last row of Table 4 restricts evaluation to the 314 TUMTraf frames whose Qwen-VL-Max description reports low illumination, overexposure, or motion blur at moderate or higher severity. The gain from an explicit description is +0.012 against +0.009 on the full set. This is what the formulation implies: the null embedding encodes the degradation distribution of the training set marginally, which is close to optimal on typical frames but mis-specified on the tail, and it is exactly on that tail that giving the text during inference has something to correct.

4.4 Ablation Studies

Table 5 evaluates the two main components across all four datasets. PFEM raises mAP@0.50:0.95 over a baseline that simply concatenates RGB and event features by 0.097, 0.040, 0.047, and 0.023 on four datasets respectively, introducing no learnable parameters. TBRM adds a further 0.043, 0.055, 0.022, and 0.021, so fusion and refinement both contribute in prior mode. The two components are complementary: PFEM accounts for the larger share of the gain on DSEC and TUMTraf, whereas TBRM contributes more on PKU-DDD17-Car, which contains a night split, consistent with text-guided refinement being most useful under degraded conditions.

Table 5: Main ablation in prior mode (mAP@0.50:0.95).

Setting	DSEC	DDD17	TUMTraf	EvKITTI
Baseline	0.354	0.412	0.474	0.462
+ PFEM	0.451	0.452	0.521	0.485
+ TBRM (full)	**0.494**	**0.507**	**0.543**	**0.506**

Table 6 examines TBRM on TUMTraf. Removing any component lowers mAP. Removing TPN (0.523) costs almost as much as the whole module (0.521, Table 5), so the text-conditioned pyramid accounts for most of the refinement gain, while spatial- (0.529) and self-attention (0.533) contribute less. Replacing confidence-weighted aggregation with NMS or Soft-NMS yields 0.537 and 0.539, respectively. Fusing the localization evidence of overlapping proposals is thus more effective than decaying their scores.

Table 6: TBRM ablation on TUMTraf. SA/SpA: self-/spatial attention; NMS and Soft-NMS replace our proposal aggregation.

Full	w/o SA	w/o SpA	w/o TPN	NMS	Soft-NMS
0.543	0.533	0.529	0.523	0.537	0.539

5 Conclusion

We presented RETDet, an RGB-Event-Text detector in which degradation-aware text acts as a detachable prior rather than a mandatory input. MLLM-generated degradation descriptions are mapped to layer-wise modulation parameters and randomly replaced by a learned null embedding during training, so textual supervision shapes the visual pathway while the network learns to operate without them. The single model thus offers two modes: prior mode needs no MLLM or text encoder and matches the cost of a purely visual detector, while conditioned mode uses explicit descriptions for further gains. Without text at inference, RETDet outperforms existing RGB-Event methods on four datasets, so the benefit of MLLM guidance is paid once during training rather than at every inference step.

AI USE STATEMENT

In preparing this manuscript, we used generative AI tools for language polishing and improving readability. We have not used generative AI tools to formulate research ideas, implement the proposed model, design experiments, or analyze experimental results. Other AI-assisted manuscript-preparation tasks do not apply to this work. Separately, as part of the proposed RETDet framework, Qwen-VL-Max was used to generate structured degradation descriptions of RGB-event image pairs, as described in Appendix A. We have reviewed all AI-assisted writing edits for technical accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The RETDet architecture, feature modulation, and proposal aggregation approach are described in Sec. 3. The text-dropout procedure and the two inference modes are specified in Sec. 3.3. Experimental hardware, backbone initialization, optimization settings, and the text-dropout probability, together with the per-dataset settings, are reported in **Appendix B.4**. Dataset descriptions and preprocessing are provided in **Appendix B.1**, while the prompt and procedure used to generate degradation descriptions are documented in **Appendix A**.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv.2308.12966*, 1, 2023.

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

Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, 2018.

Hu Cao, Guang Chen, Jiahao Xia, Genghang Zhuang, and Alois Knoll. Fusion-based feature attention gate component for vehicle detection based on event camera. *IEEE Sensors Journal*, 21(21): 24540–24548, 2021. doi: 10.1109/JSEN.2021.3115016.

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, 2024.

Guang Chen, Hu Cao, Jorg Conradt, Huajin Tang, Florian Rohrbein, and Alois Knoll. Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception. *IEEE Signal Processing Magazine*, 37(4):34–49, 2020.

Guang Chen, Kai Chen, Lijun Zhang, Liming Zhang, and Alois Knoll. Vcanet: Vanishing-point-guided context-aware network for small road object detection. *Automotive Innovation*, 4(4): 400–412, 2021.

Loïc Cordone, Benoît Miramond, and Philippe Thierion. Object detection with spiking neural networks on automotive event data. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2022. doi: 10.1109/IJCNN55064.2022.9892618.

Christian Creß, Walter Zimmer, Nils Purschke, Bach Ngoc Doan, Sven Kirchner, Venkatnarayanan Lakshminarasimhan, Leah Strand, and Alois C Knoll. Tumtraf event: Calibration and fusion resulting in a dataset for roadside event-based and rgb cameras. *IEEE Transactions on Intelligent Vehicles*, 9(7):5186–5203, 2024.

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.

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

Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021a.

Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13884–13893. IEEE, 2023.

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, 2021b.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 6469–6477. IEEE, 2017.

Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1510–1519, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Jianing Li, Siwei Dong, Zhaofei Yu, Yonghong Tian, and Tiejun Huang. Event-based vision enhanced: A joint detection framework in autonomous driving. *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1396–1401, 2019.

Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing*, 31:2975–2987, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022b.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965, 2022c.

Zichen Liang, Hu Cao, Chu Yang, Zikai Zhang, and G. Chen. Global-local feature aggregation for event-based object detection on eventkitti. *2022 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 1–7, 2022.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2017a.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017b.

Li Liu, Wanli Ouyang, Xiaogang Wang, Paul W. Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128:261 – 318, 2020.

Dongyue Lu, Lingdong Kong, Gim Hee Lee, Camille Simon Chane, and Wei Tsang Ooi. Flexevent: Towards flexible event-frame object detection at varying operational frequencies. *Advances in Neural Information Processing Systems 38*, 2025.

Yifang Ma, Zhenyu Wang, Hong Yang, and Lin Yang. Artificial intelligence applications in the development of autonomous vehicles: a survey. *IEEE/CAA Journal of Automatica Sinica*, 7: 315–329, 2020.

Diederik Paul Moeys, Federico Corradi, Chenghan Li, Simeon A. Bamford, Luca Longinotti, Fabian F. Voigt, Stewart Berry, Gemma Taverni, Fritjof Helmchen, and Tobi Delbruck. A sensitive dynamic and active pixel vision sensor for color or neural imaging applications. *IEEE Transactions on Biomedical Circuits and Systems*, 12:123–136, 2017.

Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43: 3388–3415, 2020.

Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33:16639–16652, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, 2018.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Simon Thomas Schaefer, Daniel Gehrig, and Davide Scaramuzza. Aegnn: Asynchronous event-based graph neural networks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12361–12371, 2022.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021.

Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kai-Wei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European Conference on Computer Vision*, 2022.

Abhishek Tomy, Anshul K. Paigwar, Khushdeep Singh Mann, Alessandro Renzaglia, and Christian Laugier. Fusing event-based and rgb camera for robust object detection in adverse conditions. *2022 International Conference on Robotics and Automation (ICRA)*, pp. 933–939, 2022.

Yuan-Kai Wang, Shao-En Wang, and Ping-Hsien Wu. Spike-event object detection for neuromorphic vision. *IEEE Access*, 11:5215–5230, 2023.

Ji-Jia Wu, Andy Chia-Hao Chang, Chieh-Yu Chuang, Chun-Pei Chen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Yung-Yu Chuang, and Yen-Yu Lin. Image-text co-decomposition for text-supervised semantic segmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26784–26793, 2024.

Lingxiao Yang, Ru-Yuan Zhang, Li-Da Li, and Xiaohua Xie. Simam: A simple, parameter-free attention module for convolutional neural networks. In *International Conference on Machine Learning*, 2021.

Muyang Yi, Quan Cui, Hao Wu, Cheng Yang, Osamu Yoshie, and Hongtao Lu. A simple framework for text-supervised semantic segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7071–7080, 2023.

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Rui-Ping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 24:14679–14694, 2023.

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

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

Appendix

CONTENTS

A Text Generation Details **15**

B Experiment Details **16**

B.1 Datasets 16

B.2 Evaluation Metrics 17

B.3 Loss Function 18

B.4 Implementation Details 18

C Additional Experiments and Discussions **19**

C.1 Qualitative analysis 19

C.2 Image-Based versus Text-Based Conditioning 23

D Impact and Limitations **24**

D.1 Impact Statement 24

D.2 Limitations 24

A TEXT GENERATION DETAILS

To introduce degradation-aware textual priors, this paper employs the large-scale vision-language model Qwen-VL-Max to analyze the quality of RGB and event images and generate structured semantic degradation descriptions. As a large-scale vision-language model in the Qwen-VL series (Bai et al., 2023), Qwen-VL-Max possesses strong multimodal reasoning and fine-grained visual understanding capabilities, enabling joint modeling of visual and linguistic information. Based on this, a specially designed structured prompt is used to guide the model to extract semantic degradation descriptions and generate concise responses in a fixed format, thereby providing effective support for subsequent text-guided visual feature modulation.

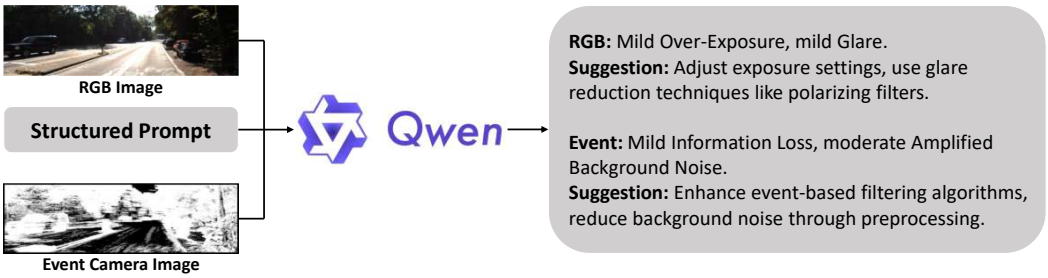


Figure 7: Example of structured degradation-aware text generated by Qwen-VL-Max for an RGB-Event image pair, including degradation descriptions and improvement suggestions.

Fig. 7 presents an example of the generated text for an RGB–event image pair. The results show that the generated text can effectively summarize the major degradation types and their severity levels in both the RGB image and the event image, while also providing corresponding suggestions for improvement, thereby offering effective semantic priors for subsequent text-guided cross-modal feature enhancement.

In our project, Qwen-VL-Max (Bai et al., 2023) is used to analyze the visual quality of RGB and event images. We design structured prompts to extract semantic degradation descriptions. These descriptions describe types such as Low-Light, Noise, or Temporal Blur, along with their severity. The model generates concise, fixed-format responses suitable for downstream encoding. These responses are then converted to feature embeddings using CLIP, enabling fusion with image features. The prompt is carefully designed to produce consistent, formal, and structured responses from the Qwen-VL-Max model. It reads as follows:

- "Analyze the first RGB image and the second event camera image separately. For each image, determine which of the following degradation types it shows, describe the severity of each degradation (e.g., mild, moderate, or severe), and briefly suggest improvement methods appropriate to each severity level. RGB degradation types: Low-Light, Over-Exposure, Haze/Fog, Glare, Noise, Blur, Color Cast, Compression Artifacts. Event camera degradation types: Information Loss, Temporal Blur, Artifacts, Polarity Bias, Amplified Background Noise, Low Spatial Resolution. Do not use formatting symbols (e.g., *, -, newlines). Use only English words, commas, and full stops. The total response must not exceed 70 words. No extra explanation. Strictly follow this format. Format Example: RGB: mild Low-Light, severe Noise. Suggestion: Use denoising, increase exposure. Event: mild Temporal Blur. Suggestion: Apply motion deblurring."

CLIP (Radford et al., 2021) uses a Transformer-based text encoder with a fixed input length. The maximum input is 77 tokens, including special tokens for start and end. These tokens are generated using BPE (Sennrich et al., 2016). On average, one English word corresponds to one or two tokens. Therefore, the practical limit is around 60 to 70 English words. Inputs longer than this are truncated. As a result, any excess information beyond the limit is discarded. This constraint ensures stable computation and efficient inference. It is important to design prompts that are concise and informative within this length.

B EXPERIMENT DETAILS

B.1 DATASETS

DSEC. DSEC (Gehrig et al., 2021b) is a large-scale event-based dataset designed for real-world driving scenarios. It provides event data with a resolution of 640 × 480 along with corresponding RGB images. The data were collected in Interlaken, Thun, and Zurich under various lighting and weather conditions. Since the event camera and RGB camera operate independently, geometric transformations based on calibration parameters are applied to align the two modalities. The dataset contains 41 training sequences and 12 test sequences. For object detection evaluation, we use the labels introduced in Tomy et al. (2022), which include cars, pedestrians, and large vehicles.

The original DSEC dataset (Gehrig et al., 2021b) consists of 53 sequences collected in three different regions of Switzerland, where the RGB images are captured by a FLIR color camera with a resolution of 1440 × 1080. Since the original DSEC dataset lacks annotations required for object detection, the experiments in this work adopt the annotated version introduced in Tomy et al. (2022). This annotated dataset contains 41 sequences in total, with 33, 3, and 5 sequences for training, validation, and testing, respectively, comprising 52,727 frames overall. The dataset covers a wide range of lighting conditions, from ideal to highly challenging, thus enabling a comprehensive evaluation of vision systems and ensuring their robustness and applicability in real-world environments. In addition, the raw event data and RGB data in DSEC are not strictly aligned, meaning that they do not correspond to frames captured at the same timestamp. To address this issue, we adopt the homographic transformation method described by Tomy et al. (2022) to map the annotations in RGB images onto the event camera frames. The transformation is formulated as follows:

$$P_{\text{event,rgb}} = K_{\text{event}} \cdot R_{\text{rgb}} \cdot R_{\text{event,rgb}} \cdot R_{\text{event}}^{T} \cdot K_{\text{rgb}}^{-1} \tag{8}$$

where K_{rgb} and K_{event} denote the intrinsic matrices of the RGB camera and the event camera, respectively; R_{rgb} and R_{event} represent the rotation matrices that transform distorted frames to undistorted frames for each camera; and $R_{\text{event,rgb}}$ denotes the rotation matrix that aligns the RGB camera coordinate system with the event camera coordinate system.

PKU-DDD17-Car. PKU-DDD17-Car (Li et al., 2019) is a large-scale event-based driving dataset collected using a DAVIS346 event camera (Binas et al., 2017). It contains 16 driving sequences recorded under diverse lighting and weather conditions. The PKU-DDD17-Car dataset covers two autonomous driving scenarios, namely daytime and nighttime, and also includes various conditions such as standard lighting, overexposure, low-light, and remote targets. The dataset provides both standard RGB frames and asynchronous event streams, making it suitable for training and evaluating models that fuse conventional frame-based and event-based modalities. The dataset has a resolution of 346 × 260 pixels, which provides sufficient detail for detecting small and fast-moving objects in real-world environments, making it particularly valuable for high-speed autonomous driving and real-time perception tasks. It has therefore been widely used for benchmarking event-based object detection and tracking algorithms. In our experiments, we use 3,154 annotated images containing only the Car category for training and evaluation.

TUMTraf. The TUMTraf dataset (Creß et al., 2024) is a large-scale multimodal dataset collected by the Technical University of Munich in real-world traffic scenarios. It contains synchronized RGB images, event data, and additional sensor information such as depth and semantic segmentation, and is specifically designed to benchmark advanced perception systems under challenging lighting and environmental conditions, including nighttime, rain, fog, and fast motion. In the TUMTraf Event dataset used in this work, both the RGB images and event images have a resolution of 640 × 480. The complete dataset contains more than 4,000 synchronized frames and a large number of 2D bounding box annotations, covering multiple object categories such as pedestrian, bicycle, car, motorcycle, bus, truck, and trailer. With high-quality multi-class annotations and synchronized multimodal data, TUMTraf is well suited for research on sensor fusion, real-time detection, and robust object tracking. In addition, it covers diverse driving scenarios such as urban intersections, rural roads, and highways, providing a reliable benchmark for object detection methods that fuse frame-based and event-based modalities. In our experiments, we use five categories from this dataset, namely car, pedestrian, bus, truck, and trailer, with a total of 2,027 annotated frames for training and evaluation.

EventKITTI. EventKITTI (Liang et al., 2022) is an event-based extension of the widely known KITTI dataset, which was originally developed for benchmarking tasks in autonomous driving, such as object detection, object tracking, and 3D reconstruction. To facilitate event-based research on this widely used benchmark, EventKITTI is constructed by synthesizing event data from the original KITTI image sequences through the Vid2e and ESIM pipelines. Specifically, Vid2e (Gehrig et al., 2020) takes consecutive image pairs as input and performs temporal interpolation to generate dense high-frame-rate sequences, thereby ensuring that the pixel displacement between adjacent frames remains below one pixel and improving the realism of synthetic motion. The interpolated images are then fed into ESIM (Rebecq et al., 2018), which simulates the response of an event camera based on intensity changes to produce temporally asynchronous event streams. Compared with conventional event datasets, EventKITTI offers advantages in both temporal and spatial resolution. Its synthetic event data preserve a spatial resolution of 1242 $\times$ 375, comparable to that of the original KITTI images. The version of EventKITTI used in this work contains 7,481 annotated frames and retains object-level annotations for cars, pedestrians, and cyclists. Therefore, EventKITTI establishes a connection between traditional frame-based datasets and event-driven vision research, enabling controlled benchmark evaluation and multimodal fusion studies in realistic autonomous driving scenarios.

Table 7: Summary of the key statistics of the four event datasets used in this work.

Dataset	Classes	Frames	Sequences	Resolution	Class Names
PKU-DDD17-Car (Li et al., 2019)	1	3,154	16	346 $\times$ 260	Car
DSEC (Gehrig et al., 2021b)	3	52,727	41	640 $\times$ 480	Car Pedestrian Large vehicle
TUMTraf (Creß et al., 2024)	5	2,027	—	640 $\times$ 480	Car Pedestrian Bus Truck Trailer
EventKITTI (Liang et al., 2022)	3	7,481	—	1242 $\times$ 375	Car Pedestrian Cyclist

B.2 Evaluation Metrics

Mean Average Precision (mAP). For object detection evaluation, the COCO benchmark uses Mean Average Precision (mAP), which averages the Average Precision (AP) over object classes and multiple Intersection over Union (IoU) thresholds (Lin et al., 2014). In addition to mAP@0.50, a more comprehensive metric mAP@0.50:0.95 averages AP across IoU thresholds from 0.5 to 0.95 with a step of 0.05:

$$mAP_{0.50:0.95} = \frac{1}{N_{classes}} \sum_{i=1}^{N_{classes}} \frac{1}{T} \sum_{IoU=0.50}^{0.95} AP_{i,IoU}. \tag{9}$$

where $N_{classes}$ is the number of classes, T is the number of IoU thresholds used for evaluation (typically 10 from 0.5 to 0.95 with a step of 0.05), and $AP_{i,IoU}$ denotes the average precision of the i-th class at a given IoU threshold.

In addition to mAP, we also report several metrics defined in the COCO evaluation protocol (Lin et al., 2014):

- AP_{50}: Average precision measured at an IoU threshold of 0.50, which reflects detection performance under a relatively loose localization requirement.
- AP_{75}: Average precision measured at an IoU threshold of 0.75, which evaluates performance under a stricter localization requirement.
- AP_S, AP_M, and AP_L: These metrics denote average precision for small, medium, and large objects, respectively. In the COCO protocol, object size is determined by pixel area.

Specifically, small objects are those with an area smaller than 32×32, medium objects have areas between 32×32 and 96×96, and large objects have areas greater than 96×96 pixels.

These metrics provide a more detailed view of the detector's behavior across different object scales and localization requirements, enabling a more comprehensive evaluation of detection performance.

Table 8: Detection performance statistics of the four datasets under COCO evaluation metrics.

Dataset	Classes	mAP@0.50:0.95	mAP@0.50	mAP@0.75	mAP_S	mAP_M	mAP_L
PKU-DDD17-Car (Li et al., 2019)	Car(all)	0.507	0.901	0.524	0.471	0.584	——
	Car(day)	0.517	0.905	0.543	0.479	0.613	——
	Car(night)	0.463	0.888	0.440	0.438	0.500	——
DSEC (Gehrig et al., 2021b)	Person	0.417	0.769	0.458	0.300	0.430	0.489
	Large Vehicle	0.432	0.820	0.477	0.316	0.446	0.500
	Car	0.632	0.903	0.731	0.535	0.654	0.640
	Average	0.494	0.831	0.555	0.384	0.510	0.543
TUMTraf (Creß et al., 2024)	Pedestrian	0.423	0.511	0.511	——	0.423	——
	Car	0.754	0.897	0.877	0.630	0.764	0.779
	Bus	0.713	0.846	0.846	——	0.779	0.503
	Truck	0.339	0.485	0.451	——	0.350	0.292
	Trailer	0.487	0.737	0.593	——	0.267	0.647
	Average	0.543	0.695	0.656	0.630	0.516	0.555
EventKITTI (Liang et al., 2022)	Car	0.667	0.905	0.818	0.548	0.670	0.749
	Pedestrian	0.360	0.734	0.317	0.129	0.410	0.504
	Cyclist	0.492	0.828	0.556	0.325	0.552	0.577
	Average	0.506	0.822	0.564	0.334	0.544	0.610

B.3 Loss Function

In our proposed framework, we adopt a hybrid loss formulation to jointly optimize both classification and regression objectives for region proposal and final detection stages. The loss begins with a binary cross-entropy formulation for classification within the Region Proposal Network (RPN). This is expressed as:

$$L_{\text{BCE,cls}}^{\text{RPN}}(p_i, p_i^*) = -\sum_i [p_i^* \log(p_i) + (1 - p_i^*) \log(1 - p_i)] \tag{10}$$

where p_i is the predicted objectness score for anchor i, and p_i^* is the corresponding ground truth label (1 for positive anchors, 0 for negative ones). This loss encourages the model to correctly distinguish object and background regions.

For bounding box regression in the RPN stage, we apply an ℓ_1 loss that measures the difference between predicted and target box coordinates. This is weighted by the objectness label to focus only on positive samples:

$$L_{l_1,\text{reg}}^{\text{RPN}}(p_i^*, t_i, t_i^*) = \sum_i p_i^* \cdot R(t_i - t_i^*) \tag{11}$$

Here, t_i represents the predicted box offsets, and t_i^* the ground truth offsets. The function $R(\cdot)$ denotes the smooth ℓ_1 loss used to handle outliers more robustly than traditional squared error.

Finally, the overall training objective L_{total} combines the classification and regression losses from the RGB/event RPN branches with the final classification and bounding-box regression losses from the detection head:

$$L_{\text{total}} = L_{\text{BCE,cls}}^{\text{RPN,RGB}} + L_{\text{BCE,cls}}^{\text{RPN,Event}} + L_{\text{L}_1,\text{reg}}^{\text{RPN,RGB}} + L_{\text{L}_1,\text{reg}}^{\text{RPN,Event}} + L_{\text{CE,cls}}^{\text{Det}} + L_{\text{L}_1,\text{reg}}^{\text{Det}} \tag{12}$$

B.4 Implementation Details

All experiments are conducted on Ubuntu 20.04 with a single NVIDIA RTX 3090 GPU (24GB), and the implementation is based on PyTorch. Following the training settings of Tomy et al. (2022), all models are trained with the Adam optimizer (Kingma & Ba, 2015). The CLIP text encoder uses the

ViT-B/32 backbone, where 'B' denotes the base-scale model and '32' the patch size of the visual encoder. ResNet and RVT backbones are initialized with pretrained weights. During the first three epochs the backbone is frozen and the remaining modules are trained with a learning rate of 1×10^{-4}; after unfreezing, the backbone learning rate is set to 1×10^{-5} while the other modules remain at 1×10^{-4}. Text dropout is applied with probability $p = 0.15$ throughout training. The per-dataset settings are as follows.

- For the DSEC dataset, which has a resolution of 640 × 480, a batch size of 4 is used during training. Both the FPN and TPN adopt a five-level structure, including P2, P3, P4, P5, and P6. In ROI Align, the parameter k_0 is set to 4 and the scale parameter s is set to 224. The detection model is trained for 60 epochs on this dataset.
- For the PKU-DDD17-Car dataset, since its resolution is 346 × 260, a relatively large batch size of 8 is adopted during training. Specifically, ResNet-50 and RVT are used as backbone networks to extract feature maps $C2, C3, C4, C5$, which are then fed into a four-layer FPN to generate $P2, P3, P4, P5$. The TPN is configured in the same way. In ROI Align, the parameter k_0 is set to 3 and the scale parameter s is set to 112. The detection model is trained for 120 epochs on this dataset.
- For the TUMTraf dataset, which has the same resolution of 640 × 480, a batch size of 4 is also adopted during training. Similar to DSEC, both the FPN and TPN use a five-layer structure. In ROI Align, the parameter k_0 is set to 3 and the scale parameter s is set to 224. The detection model is trained for 60 epochs on this dataset.
- For the EventKITTI dataset, since its image resolution is 1242 × 375, a batch size of 2 is adopted during training. Under this setting, both the FPN and TPN use a four-layer structure. In ROI Align, the parameter k_0 is set to 4 and the scale parameter s is set to 224. The detection model is trained for 60 epochs on this dataset.

C Additional Experiments and Discussions

C.1 Qualitative analysis

To further validate the proposed method, Fig. 8, Fig. 9, Fig. 10, and Fig. 11 present qualitative comparisons with CAFR on representative scenes from different datasets. Our method achieves higher detection sensitivity in complex environments, especially in scenes with dense objects or distant small targets. For example, in Scene 5 of Fig. 8, our method obtains more complete detection results under low-light conditions and accurately identifies potential targets missing from the ground truth. In Scene 4 of Fig. 11, our method better detects small-scale vehicles under overexposed degradation. These results demonstrate that, compared with relying solely on RGB-Event visual feature fusion, introducing the textual modality provides additional semantic guidance and enhances feature representation, thereby further improving object detection performance in complex scenarios.

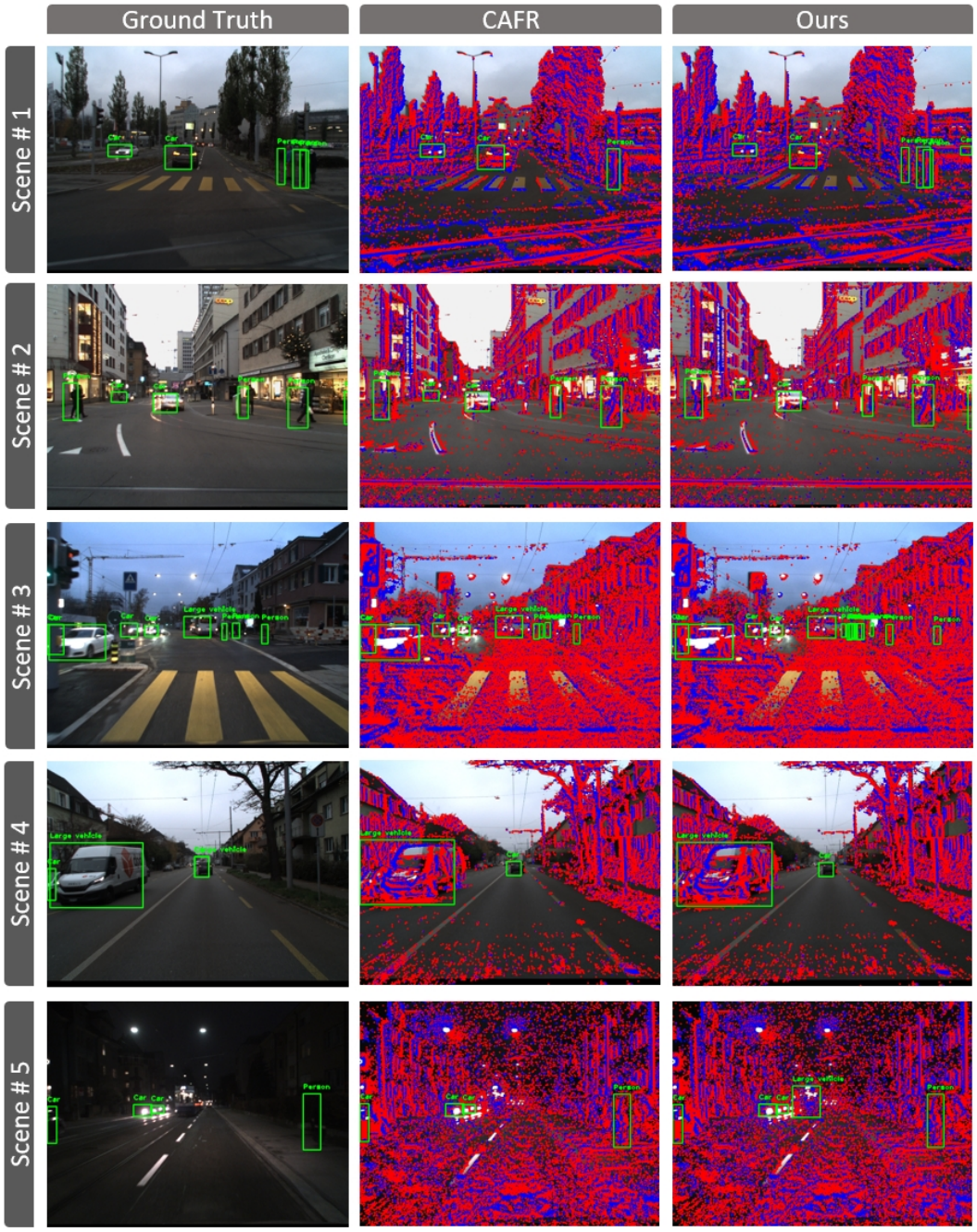


Figure 8: Qualitative comparison of object detection results between our method and CAFR (Cao et al., 2024) across multiple scenes on the DSEC dataset. It can be seen that our method achieves better detection performance in challenging scenarios with complex backgrounds, low-light conditions, and dense targets, especially showing higher accuracy in pedestrian detection.

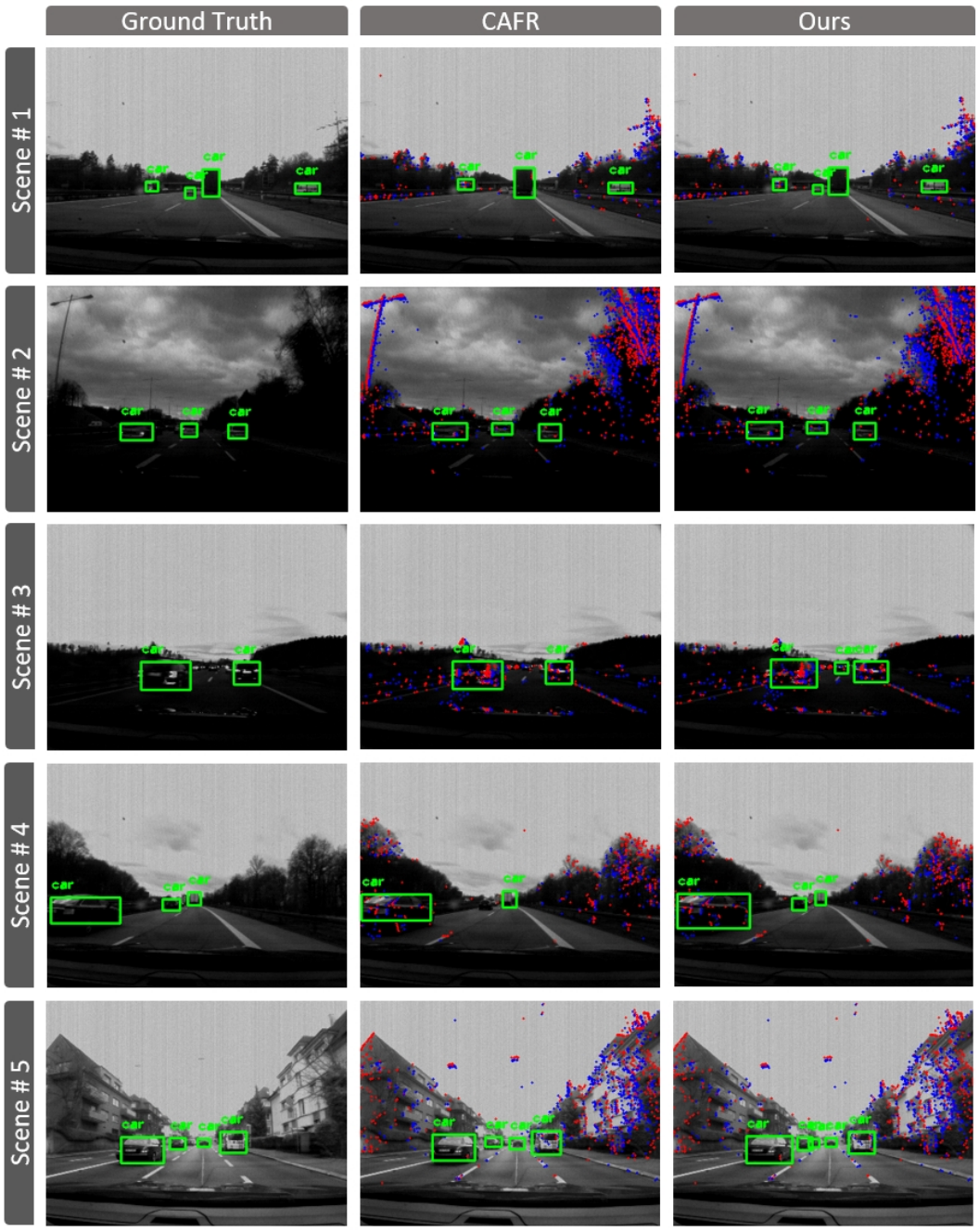


Figure 9: Qualitative comparison of object detection results between our method and CAFR (Cao et al., 2024) across multiple scenes on the PKU-DDD17-Car dataset. From left to right are the ground truth, the detection results of CAFR, and the detection results of our method. It can be observed that our method achieves better detection performance across multiple scenes, especially for small-scale objects.

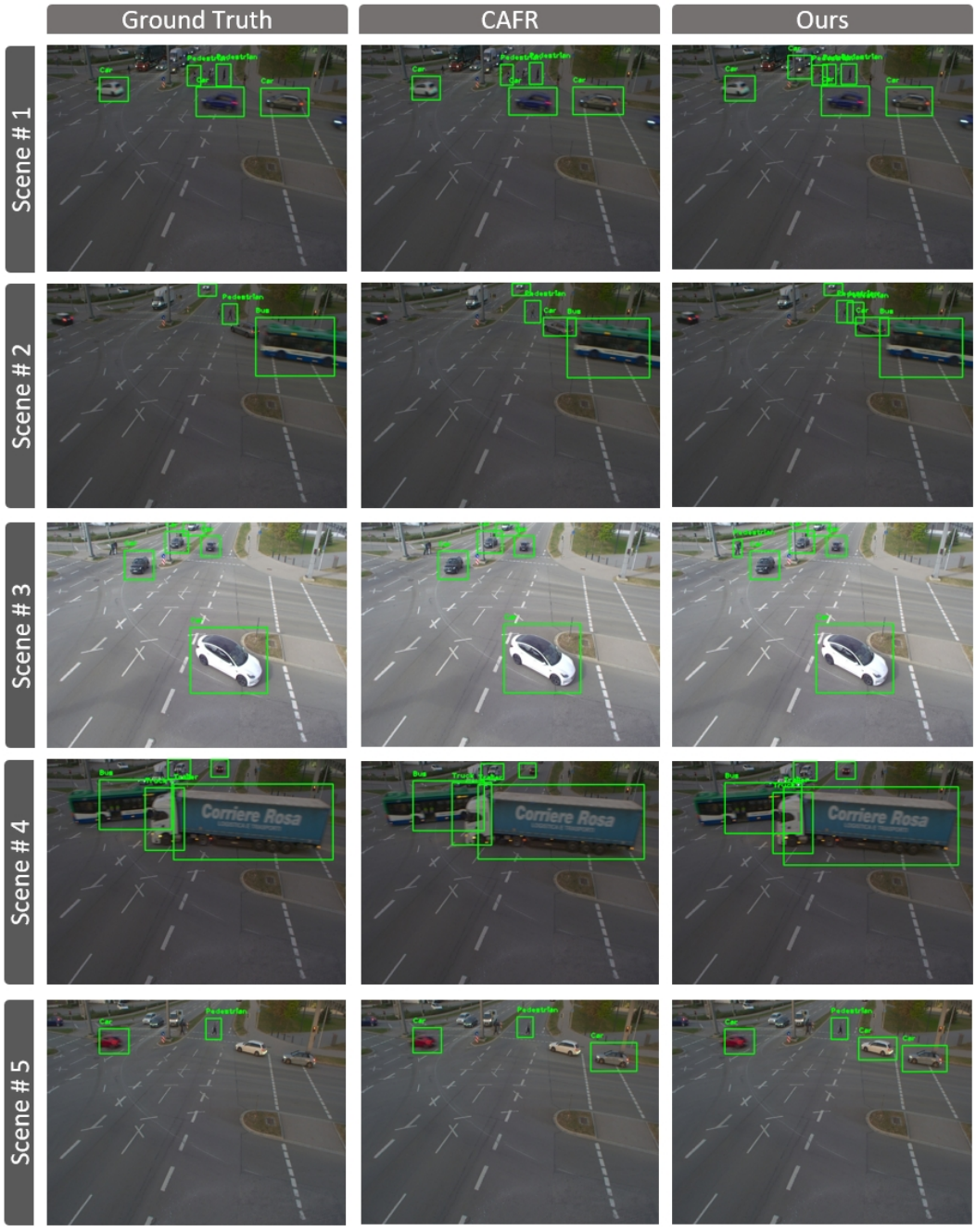


Figure 10: Qualitative comparison of object detection results between our method and CAFR (Cao et al., 2024) across multiple scenes on the TUMTraf dataset. Our method not only improves small-scale object detection under complex traffic conditions, but also accurately detects potential targets that are not included in the ground-truth annotations.

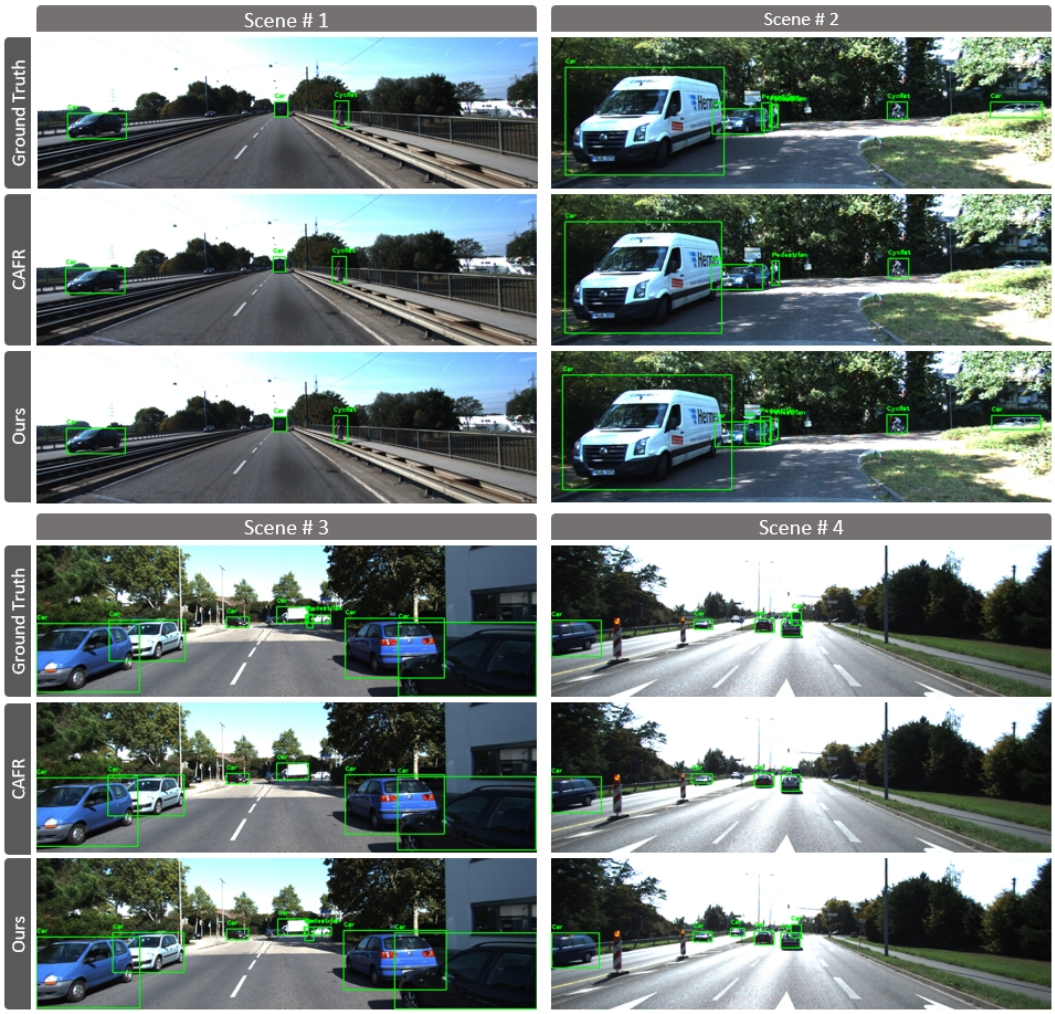


Figure 11: Qualitative comparison of object detection results between our method and CAFR (Cao et al., 2024) across multiple scenes on the EventKITTI dataset. It can be observed that, in some scenes, our method still achieves better detection performance for distant small-scale objects.

C.2 Image-Based versus Text-Based Conditioning

To examine whether the conditioning signal can be derived directly from the input image, we replace the text embedding with a CLIP image embedding while keeping the remaining components unchanged. As shown in Table 9, this variant reaches 0.513 mAP on TUMTraf, below both prior and text-conditioned modes. We read this as redundancy: the embedding is a deterministic function of a frame the detector already processes, so it supplies no knowledge the visual pathway lacks while adding a shortcut the training objective can exploit. A description is produced by a larger external model and compresses the scene into a judgement about sensing conditions, and it is that compression that makes it a prior rather than a second copy of the input.

Table 9: Comparison of conditioning sources on TUMTraf. All results are reported as mAP@0.50:0.95.

Conditioning source	mAP@0.50:0.95
Learned null embedding (prior mode)	0.543
CLIP image embedding	0.513
CLIP text embedding (conditioned mode)	**0.552**

D IMPACT AND LIMITATIONS

D.1 IMPACT STATEMENT

Our method has the potential to benefit safety-critical perception systems, especially in autonomous driving and intelligent transportation. By combining RGB images, event streams, and degradation-aware textual priors, the proposed framework aims to improve object detection under challenging conditions such as low illumination, motion blur, rapid motion, and highly dynamic scenes, which may help detect vehicles, pedestrians, cyclists, and other road participants more reliably. Its efficiency-oriented design, including the parameter-free PFEM and the proposal refinement strategy in TBRM, also makes it more suitable for real-world systems with limited computational resources. More broadly, this work may contribute to multimodal perception research by exploring the integration of event-based sensing and text-guided semantic information, and may be useful for related applications such as roadside traffic monitoring, mobile robotics, drone navigation, infrastructure inspection, and intelligent monitoring. However, stronger perception capabilities may also have potential negative impacts. If this technology is applied to surveillance systems, city-scale sensing networks, or large-scale public-space observation platforms, it may intensify concerns regarding personal privacy, the boundaries of data collection, and the ways in which information is used. For example, in the absence of sufficient notification, explicit authorization, or effective regulation, continuous visual perception systems could be misused for inappropriate tracking, behavior analysis, or crowd monitoring, thereby affecting individuals' rights to privacy and freedom of movement. In addition, although the introduction of the text modality can provide semantic priors, it may also introduce new risks. For example, biases, inaccuracies, or incomplete information in textual descriptions may be inherited or even amplified by the model, thereby affecting the fairness, reliability, and interpretability of perception results.

D.2 LIMITATIONS

RETDet relies on MLLM-generated degradation descriptions during training. Inaccurate or incomplete descriptions may limit the learned prior, even though prior mode requires no textual input at inference. Conditioned mode additionally depends on descriptions matching the current scene; cached or preset descriptions may become less informative as sensing conditions change.

Removing inference-time text generation does not eliminate the detector's computational cost. Prior mode takes 306.6 ms per image in the reported TUMTraf setting, which remains a constraint for real-time deployment. Online description generation adds over one second per image in conditioned mode. Model compression and system optimization are therefore needed for latency-sensitive or resource-constrained applications.

The Bayesian interpretation of proposal aggregation assumes conditionally independent Gaussian observations and uses objectness as a proxy for inverse variance. Overlapping proposals can be correlated, and objectness is not calibrated localization uncertainty. The resulting estimator is therefore an approximation whose assumptions may be less suitable in crowded or ambiguous scenes.